Docket No.:PF-0232-1 DIV

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Box Non-Fee Amendment, Commissioner for Patents, Washington, D.C. 20231 on December 13, 2002.

By: ______________ Printed: Katherine Stofer

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: LaBrie et al.

Title: HUMAN TUBBY HOMOLOG

Serial No.: 09/782,390 Filing Date: February 12, 2001

Examiner: Spector, L. Group Art Unit: 1647

Box Non-Fee Amendment
Commissioner for Patents
Washington, D.C. 20231

DECLARATION OF LARS MICHAEL FURNESS UNDER 37 C.F.R. § 1.132

I, L. MICHAEL FURNESS, a citizen of the United Kingdom, residing at 2 Brookside, Exning, Newmarket, United Kingdom, declare that:

1. I was employed by Incyte Genomics, Inc. (hereinafter "Incyte") as a Director of Pharmacogenomics until December 31, 2001. I am currently under contract to be a Consultant to Incyte Genomics, Inc.

2. In 1984, I received a B.Sc.(Hons) in Biomolecular Science (Biophysics and Biochemistry) from Portsmouth Polytechnic.

From 1985-1987 I was at the School of Pharmacy in London, United Kingdom, during which time I analyzed lipid methyltransferase enzymes using a variety of protein analysis methods, including one-dimensional (1D) and two-dimensional (2D) gel electrophoresis, HPLC, and a variety of enzymatic assay systems.

I then worked in the Protein Structure group at the National Institute for Medical Research until 1989, setting up core facilities for nucleic acid synthesis and sequencing, as well as assisting in programs on protein kinase C inhibitors.

After a year at Perkin Elmer-Applied Biosystems as a technical specialist, I worked at the Imperial Cancer Research Fund between 1990-1992, on a Eureka-funded program collaborating with Amersham Pharmacia in the United Kingdom and CEPH (Centre d'Etude du Polymorphisme Humaine) in Paris, France, to develop novel nucleic acid purification and characterization methods.

In 1992, I moved to Pfizer Central Research in the United Kingdom, where I stayed until 1998, initially setting up core DNA sequencing and then a DNA arraying facility for gene expression analysis in 1993. My work also included bioinformatics and I was responsible for the support of all Pfizer neuroscience programs in the United Kingdom. This then led me into carrying out detailed bioinformatics and wet lab work on the sodium channels, including antibody generation, Western and Northern analyses, PCR, tissue distribution studies, and sequence analyses on novel sequences identified.

In 1998, I moved to Incyte Genomics, Inc., to the Pharmacogenomics group, to look at the application of genomics and proteomics to the pharmaceutical industry. In 1999, I was appointed director of the LifeExpress Lead Program which used microarray and protein expression data to identify pharmacologically and toxicologically relevant mechanisms to assist in improved drug design and development.

On December 12, 2001, I founded Nuomics Consulting Ltd., in Exning, U.K., and I am currently employed as Managing Director. Nuomics Consulting Ltd. will be providing expert technical knowledge and advice to businesses around the areas of genomics, proteomics, pharmacogenomics, toxicogenomics and chemogenomics.

3. I have reviewed the specification of a United States patent application that I understand was filed on February 12, 2001 in the names of Samuel T. LaBrie et al., and was assigned Serial No. 09/782,390 (hereinafter "the LaBrie '390 application"). Furthermore, I understand that this United States patent application was a divisional application of and claimed

priority to United States patent application Serial No. 08/812,824 filed on March 6, 1997 (hereinafter "the LaBrie '824 application"), having essentially the identical specification, with the exception of corrected typographical errors and reformatting changes. Thus page and line numbers may not match as between the LaBrie '390 application and the LaBrie '824 application. My remarks herein will therefore be directed to the LaBrie '824 patent application, and March 6, 1997, as the relevant date of filing. In broad overview, the LaBrie '824 specification pertains to certain nucleotide and amino acid sequences and their use in a number of applications, including gene and protein expression monitoring applications that are useful in connection with (a) developing drugs (e.g., for the treatment of cancer), and (b) monitoring the activity of drugs for purposes relating to evaluating their efficacy and toxicity.

4. I understand that (a) the LaBrie '390 application contains claims that are directed to a substantially purified polypeptide having the sequence shown as SEQ ID NO:1 (hereinafter "the SEQ ID NO:1 polypeptide"), and (b) the Patent Examiner has rejected those claims on the grounds that the specification of the LaBrie '390 application does not disclose a substantial, specific and credible utility for the claimed SEQ ID NO:1 polypeptide. I further understand that whether or not a patent specification discloses a substantial, specific and credible utility for its claimed subject matter is properly determined from the perspective of a person skilled in the art to which the specification pertains at the time of the patent application was filed. In addition, I understand that a substantial, specific and credible utility under the patent laws must be a "real-world" utility.

5. I have been asked (a) to consider with a view to reaching a conclusion (or conclusions) as to whether or not I agree with the Patent Examiner's position that the LaBrie '390 application and its parent, the LaBrie '824 application, does not disclose a substantial, specific and credible "real-world" utility for the claimed SEQ ID NO:1 polypeptide, and (b) to state and explain the bases for any conclusions I reach. I have been informed that, in connection with my considerations, I should determine whether or not a person skilled in the art to which the LaBrie '824 application pertains on March 6, 1997, would have concluded that the LaBrie '824

application disclosed, for the benefit of the public, a specific beneficial use of the SEQ ID NO:1 polypeptide in its then available and disclosed form. I have also been informed that, with respect to the "real-world" utility requirement, the Patent and Trademark Office instructs its Patent Examiners in Section 2107 of the Manual of Patent Examining Procedure, under the heading "I. 'Real-World Value' Requirement":

> "Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact 'useful' in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm."

6. I have considered the matters set forth in paragraph 5 of this Declaration and have concluded that, contrary to the position I understand the Patent Examiner has taken, the specification of the LaBrie '824 patent application disclosed to a person skilled in the art at the time of its filing a number of substantial, specific and credible real-world utilities for the claimed SEQ ID NO:1 polypeptide. More specifically, persons skilled in the art on March 6, 1997 would have understood the LaBrie '824 application to disclose the use of the SEQ ID NO:1 polypeptide as a research tool in a number of gene and protein expression monitoring applications that were well-known at that time to be useful in connection with the development of drugs and the monitoring of the activity of such drugs. I explain the bases for reaching my conclusion in this regard in paragraphs 7-13 below.

7. In reaching the conclusion stated in paragraph 6 of this Declaration, I considered (a) the specification of the LaBrie '824 application, and (b) a number of published articles and patent documents that evidence gene and protein expression monitoring techniques that were well-known before the March 6, 1997 filing date of the LaBrie '824 application. The published articles and patent documents I considered are:

(a) Anderson, N.L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N.G., A Two-Dimensional Gel Database of Rat Liver Proteins Useful in Gene Regulation and Drug Effects Studies, Electrophoresis, 12, 907-930 (1991) (hereinafter "the Anderson 1991 article") (copy annexed at Tab A);

(b) Anderson, N.L., Esquer-Blasco, R., Hofmann, J.-P., Mehues, L., Raymackers, J., Steiner, S. Witzmann, F., Anderson, N.G., An Updated Two-Dimensional Gel Database of Rat Liver Proteins Useful in Gene Regulation and Drug Effect Studies, Electrophoresis, 16, 1977-1981 (1995) (hereinafter "the Anderson 1995 article") (copy annexed at Tab B);

(c) Wilkins, M.R., Sanchez, J.-C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., Williams, K.L., Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It, Biotechnology and Genetic Engineering Reviews, 13, 19-50 (1995) (hereinafter "the Wilkins article") (copy annexed at Tab C);

(d) Celis, J.E., Rasmussen, H.H., Leffers, H., Madsen, P., Honore, B., Gesser, B., Dejgaard, K., Vandekerckhove, J., Human Cellular Protein Patterns and their Link to Genome DNA Sequence Data: Usefulness of Two-Dimensional Gel Electrophoresis and Microsequencing, FASEB Journal, 5, 2200-2208 (1991) (hereinafter "the Celis article") (copy annexed at Tab D);

(e) Franzen, B., Linder, S., Okuzawa, K., Kato, H., Auer, G., Nonenzymatic Extraction of Cells from Clinical Tumor Material for Analysis of Gene Expression by Two-Dimensional Polyacrylamide Gel Electrophoresis, Electrophoresis, 14, 1045-1053 (1993) (hereinafter "the Franzen article") (copy annexed at Tab E);

(f) Bjellqvist, B., Basse, B., Olsen, E., Celis, J.E., Reference Points for Comparisons of Two-Dimensional Maps of Proteins from Different Human Cell Types Defined in a pH Scale Where Isoelectric Points Correlate with Polypeptide Compositions, Electrophoresis, 15, 529-539 (1994) (hereinafter "the Bjellqvist article") (copy annexed at Tab F);

(g) Large Scale Biology Company Info; LSB and LSP Information;

from http://www.lsbc.com (2001) (copy annexed at Tab G);

8. Many of the published articles I considered (i.e., at least items (a)-(f) identified in paragraph 7) relate to the development of protein two-dimensional gel electrophoretic techniques for use in gene expression monitoring applications in drug development and toxicology. As I will discuss below, a person skilled in the art who read the LaBrie '824 application on March 6, 1997 would have understood that application to disclose the SEQ ID NO:1 polypeptide to be useful for a number of gene and protein expression monitoring applications, e.g., in the use of two-dimensional polyacrylamide gel electrophoresis and western blot analysis of tissue samples in drug development and in toxicity testing.

9. Turning more specifically to the LaBrie '824 specification, the SEQ ID NO:1 polypeptide is shown at pages 47-49 as one of four sequences under the heading "Sequence Listing." The LaBrie '824 specification specifically teaches that the "invention features a novel human tubby homolog (NHT) having the amino acid sequence shown in SEQ ID NO:1" (LaBrie '824 application at p. 2). It further teaches that (a) the identity of the SEQ ID NO:1 polypeptide was determined from a "neuronal cell cDNA library", (b) the SEQ ID NO:1 polypeptide is the novel tubby homolog referred to as "NHT" and is encoded by SEQ ID NO:2, and (c) northern analysis shows that "NHT is expressed in brain and neuronal tissues and lymph node tissue," and therefore "NHT appears to be involved in mammalian appetite and eating disorders, and to play a role in appetite and eating disorders, especially anorexia, cachexia and obesity" (LaBrie '824 application at p. 10, line 30 to p. 31, line15, p. 23, lines 11-13, and p. 2, lines 12-15).

The LaBrie '824 application discusses a number of uses of the SEQ ID NO:1 polypeptide in addition to its use in gene expression monitoring applications. I have not fully evaluated these additional uses in connection with the preparation of this Declaration and do not express any views in this Declaration regarding whether or not the LaBrie '824 specification discloses these additional uses to be substantial, specific and credible real-world utilities of the SEQ ID NO:1 polypeptide. Consequently, my discussion in this Declaration concerning the LaBrie '824 application focuses on the portions of the application that relate to the use of the

SEQ ID NO:1 polypeptide in gene and protein expression monitoring applications.

10. The LaBrie '824 application discloses that the polynucleotide sequences disclosed therein, including the polynucleotides encoding the SEQ ID NO:1 polypeptide, are useful as probes in chip based technologies. It further teaches that the chip based technologies can be used "for the detection and/or quantification of nucleic acid or protein" (LaBrie '824 application at p. 21, lines8-10).

The LaBrie '824 application also discloses that the SEQ ID NO:1 polypeptide is useful in other protein expression detection technologies. The LaBrie '824 application states that "[a] variety of protocols for detecting and measuring the expression of NHT, using either polyclonal or monoclonal antibodies specific for the protein are known in the art. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), and fluorescence activated cell sorting (FACS)" (LaBrie '824 application at p. 21, lines 19-22). Furthermore, the LaBrie '824 application discloses that "[a] variety of protocols including ELISA, RIA, and FACS for measuring NHT are known in the art and provide a basis for diagnosing altered or abnormal levels of NHT expression. Normal or standard values for NHT expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably human, with antibody to NHT under conditions suitable for complex formation" (LaBrie '824 application at p.32, lines 24-28).

In addition, at the time of filing the LaBrie '824 application, it was well known in the art that "gene" and protein expression analyses also included two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) technologies, which were developed during the 1980s, and as exemplified by the Anderson 1991 and 1995 articles (Tab A and Tab B). The Anderson 1991 article teaches that a 2-D PAGE map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies including regulation of protein expression by various drugs and toxic agents (Tab A at p. 907). The Anderson 1991 article teaches an empirically-determined standard curve fitted to a series of identified proteins based upon amino acid chain length (Tab A at p. 911) and how that standard curve can be used in protein expression analysis. The Anderson 1991 article teaches that "there is a long-term need

for a comprehensive database of liver proteins" (Tab A at p. 912).

The Wilkins article is one of a number of documents that were published prior to the March 6, 1997 filing date of the LaBrie '824 application that describes the use of the 2-D PAGE technology in a wide range of gene and protein expression monitoring applications, including monitoring and analyzing protein expression patterns in human cancer, human serum plasma proteins, and in rodent liver following exposure to toxins. In view of the LaBrie '824 application, the Wilkins article, and other related pre-March 6, 1997 publications, persons skilled in the art on March 6, 1997 clearly would have understood the LaBrie '824 application to disclose the SEQ ID NO:1 polypeptide to be useful in 2-D PAGE analyses for the development of new drugs and monitoring the activities of drugs for such purposes as evaluating their efficacy and toxicity, as explained more fully in paragraph 12 below.

With specific reference to toxicity evaluations, those of skill in the art who were working on drug development in March 1997 (and for many years prior to March 1997) without any doubt appreciated that the toxicity (or lack of toxicity) of any proposed drug they were working on was one of the most important criteria to be considered and evaluated in connection with the development of the drug. They would have understood at that time that good drugs are not only potent, they are specific. This means that they have strong effects on a specific biological target and minimal effects on all other biological targets. Ascertaining that a candidate drug affects its intended target, and identification of undesirable secondary effects (i.e., toxic side effects), had been for many years among the main challenges in developing new drugs. The ability to determine which genes are positively affected by a given drug, coupled with the ability to quickly and at the earliest time possible in the drug development process identify drugs that are likely to be toxic because of their undesirable secondary effects, have enormous value in improving the efficiency of the drug discovery process, and are an important and essential part of the development of any new drug. In fact, the desire to identify and understand toxicological effects using the experimental assays described above led Dr Leigh Anderson to found the Large Scale Biology Corporation in 1985, in order to pursue commercial development of the 2-D electrophoretic protein mapping technology he had developed. In addition, the company focused on toxicological effects on the proteome as clearly demonstrated by its goals and by its senior

management credentials described in company documents (see Tab G at pp. 1, 3, and 5).

Accordingly, the teachings in the LaBrie '824 application, in particular regarding use of SEQ ID NO:1 in differential gene and protein expression analysis (2-D PAGE maps) and in the development and the monitoring of the activities of drugs, clearly includes toxicity studies and persons skilled in the art who read the LaBrie '824 application on March 6, 1997 would have understood that to be so.

11. As previously discussed (*supra*, paragraphs 7 and 8), my experience with protein analysis methods in the mid-1980s and the several publications annexed to this Declaration at Tabs A through F evidence information that was available to the public regarding two-dimensional polyacrylamide gel electrophoresis technology and its uses in drug discovery and toxicology testing before the March 6, 1997 filing date of the LaBrie '824 application. In particular the Celis article stated that "protein databases are expected to foster a variety of biological information.... -- among others, ... drug development and testing" (See Tab D, p. 2200, second column). The Franzen article shows that 2-D PAGE maps were used to identify proteins in clinical tumor material (See Tab E). The LaBrie '824 application clearly discloses that expression of NHT is associated with brain, neuronal and lymph node tissues (LaBrie '824 application at p. 11, lines 13-15). The Bjellqvist article showed that a protein may be identified accurately by its positional co-ordinates, namely molecular mass and isoelectric point (See Tab F). The LaBrie '824 application clearly disclosed SEQ ID NO:1 from which it would have been routine for one of skill in the art to predict both the molecular mass and the isoelectric point using algorithms well known in the art at the time of filing.

12. A person skilled in the art on March 6, 1997, who read the LaBrie '824 application, would understand that application to disclose the SEQ ID NO:1 polypeptide to be highly useful in analysis of differential expression of proteins. For example, the specification of the LaBrie '824 application would have led a person skilled in the art in March 1997 who was using protein expression monitoring in connection with working on developing new drugs for the treatment of an appetite and eating disorders, especially anorexia, cachexia and obesity to

conclude that a 2-D PAGE map that used the substantially purified SEQ ID NO:1 polypeptide would be a highly useful tool and to request specifically that any 2-D PAGE map that was being used for such purposes utilize the SEQ ID NO:1 polypeptide sequence. Expressed proteins are useful for 2-D PAGE analysis in toxicology expression studies for a variety of reasons, particularly for purposes relating to providing controls for the 2-D PAGE analysis, and for identifying sequence or post-translational variants of the expressed sequences in response to exogenous compounds. Persons skilled in the art would appreciate that a 2-D PAGE map that utilized the SEQ ID NO:1 polypeptide sequence would be a more useful tool than a 2-D PAGE map that did not utilize this protein sequence in connection with conducting protein expression monitoring studies on proposed (or actual) drugs for treating appetite and eating disorders, especially anorexia, cachexia and obesity for such purposes as evaluating their efficacy and toxicity.

I discuss in more detail in items (a)-(b) below a number of reasons why a person skilled in the art, who read the LaBrie '824 specification in March 1997, would have concluded based on that specification and the state of the art at that time, that SEQ ID NO:1 polypeptide would be a highly useful tool for analysis of a 2-D PAGE map for evaluating the efficacy and toxicity of proposed drugs for appetite and eating disorders, especially anorexia, cachexia and obesity by means of 2-D PAGE maps, as well as for other evaluations:

(a) The LaBrie '824 specification contains a number of teachings that would lead persons skilled in the art on March 6, 1997 to conclude that a 2-D PAGE map that utilized the substantially purified SEQ ID NO:1 polypeptide would be a more useful tool for gene expression monitoring applications relating to drugs for treating appetite and eating disorders, especially anorexia, cachexia and obesity than a 2-D PAGE map that did not use the SEQ ID NO:1 polypeptide sequence. Among other things, the LaBrie '824 specification teaches that (i) the identity of the SEQ ID NO:1 polypeptide was determined from a "neuronal cell line cDNA library," (ii) the SEQ ID NO:1 polypeptide is the novel Tubby homolog referred to as NHT, and (iii) NHT is expressed in various libraries derived from brain and neuronal tissue (fetal and infant brain) and lymph node tissues and, therefore, "NHT appears to be involved in maturity onset diabetes, insulin resistance, progressive retinal degeneration and hearing loss, and to play a

role in appetite and eating disorders, especially anorexia, cachexia and obesity" (LaBrie '824 application at pp. 2; see paragraph 9, *supra*). The substantially purified polypeptide could therefore be used as a control to more accurately gauge the expression of NHT in the sample and consequently more accurately gauge the effect of a toxicant on expression of the gene.

(b) Persons skilled in the art on March 6, 1997 would have appreciated (i) that the protein expression monitoring results obtained using a 2-D PAGE map that utilized a SEQ ID NO:1 polypeptide would vary, depending on the particular drug being evaluated, and (ii) that such varying results would occur both with respect to the results obtained from the SEQ ID NO:1 polypeptide and from the 2-D PAGE map as a whole (including all its other individual proteins). These kinds of varying results, depending on the identity of the drug being tested, in no way detracts from my conclusion that persons skilled in the art on March 6, 1997, having read the LaBrie '824 specification, would specifically request that any 2-D PAGE map that was being used for conducting protein expression monitoring studies on drugs for treating appetite and eating disorders, especially anorexia, cachexia and obesity (*e.g.*, a toxicology study or any efficacy study of the type that typically takes place in connection with the development of a drug) utilize the SEQ ID NO:1 polypeptide sequence. Persons skilled in the art on March 6, 1997 would have wanted their 2-D PAGE map to utilize the SEQ ID NO:1 polypeptide sequence because a 2-D PAGE map that utilized protein sequence information the polypeptide (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies using 2-D PAGE maps that persons skilled in the art have been doing since well prior to March 6, 1997.

The foregoing is not intended to be an all-inclusive explanation of all my reasons for reaching the conclusions stated in this paragraph 12, and in paragraph 6, *supra*. In my view, however, it provides more than sufficient reasons to justify my conclusions stated in paragraph 6 of this Declaration regarding the LaBrie '824 application disclosing to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the SEQ ID NO:1 polypeptide.

13. Also pertinent to my considerations underlying this Declaration is the fact

that the LaBrie '824 disclosure regarding the uses of the SEQ ID NO:1 polypeptide for protein expression monitoring applications is not limited to the use of that protein in 2-D PAGE maps. For one thing, the LaBrie '824 disclosure regarding the technique used in gene and protein expression monitoring applications is broad (LaBrie '824 application at, e.g., p. 21, lines 6-10 and p. 32, line 24 to p. 33, line 2).

In addition, the LaBrie '824 specification repeatedly teaches that the protein described therein (including the SEQ ID NO:1 polypeptide) may desirably be used in any of a number of long established "standard" techniques, such as ELISA or western blot analysis, for conducting protein expression monitoring studies. See, e.g.:

(a) LaBrie '824 application at [p. 21, lines 19-22 ("A variety of protocols for detecting and measuring the expression of NHT, using either polyclonal or monoclonal antibodies specific for the protein are known in the art. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), and fluorescence activated cell sorting (FACS)");

(b) LaBrie '824 application at p. 32, line 24 to p. 33, line 2 ("A variety of protocols including ELISA, RIA, and FACS for measuring NHT are known in the art and provide a basis for diagnosing altered or abnormal levels of NHT expression. Normal or standard values for NHT expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably human, with antibody to NHT under conditions suitable for complex formation. The amount of standard complex formation may be quantified by various methods, but preferably by photometric, means. Quantities of NHT expressed in subject, control and disease, samples from biopsied tissues are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease").

Thus a person skilled in the art on March 6, 1997, who read the LaBrie '824 specification, would have routinely and readily appreciated that the SEQ ID NO:1 polypeptide disclosed therein would be useful to conduct gene expression monitoring analyses using 2-D PAGE mapping or western blot analysis or any of the other traditional membrane-based protein expression monitoring techniques that were known and in common use many years prior to the

filing of the LaBrie '824 application. For example, a person skilled in the art in March 1997 would have routinely and readily appreciated that the SEQ ID NO:1 polypeptide would be a useful tool in conducting protein expression analyses, using the 2-D PAGE mapping or western analysis techniques, in furtherance of (a) the development of drugs for the treatment of appetite and eating disorders, especially anorexia, cachexia and obesity, and (b) analyses of the efficacy and toxicity of such drugs.

14. I declare further that all statements made herein of my own knowledge are true and that all statements made herein on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, and that willful false statements may jeopardize the validity of this application and any patent issuing thereon.

L. Michael Furness, B.Sc.

Signed at Exning, United Kingdom
this 12th day of December, 2002

N. Leigh Anderson
Ricardo Esquer-Blasco
Jean-Paul Hofmann
Norman G. Anderson

Large Scale Biology Corporation, Rockville, MD

A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies

A standard two-dimensional (2-D) protein map of Fischer 344 rat liver (F344MST3) is presented, with a tabular listing of more than 1200 protein species. Sodium dodecyl sulfate (SDS) molecular mass and isoelectric point have been established, based on positions of numerous internal standards. This map has been used to connect and compare hundreds of 2-D gels of rat liver samples from a variety of studies, and forms the nucleus of an expanding database describing rat liver proteins and their regulation by various drugs and toxic agents. An example of such a study, involving regulation of cholesterol synthesis by cholesterol-lowering drugs and a high-cholesterol diet, is presented. Since the map has been obtained with a widely used and highly reproducible 2-D gel system (the Iso-Dalt® system), it can be directly related to an expanding body of work in other laboratories.

Contents

Correspondence: Dr. N. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850, USA

Abbreviations: CBB, Coomassie Brilliant Blue; CPK, creatine phosphokinase; 2-D, two-dimensional; IEF, isoelectric focusing; MSN, master spot number; NP-40, Nonidet P-40, SDS, sodium dodecyl sulfate

1 Introduction

High-resolution two-dimensional electrophoresis of proteins, introduced in 1975 by O'Farrell and others [1–4], has been used over the ensuing 16 years to examine a wide variety of biological systems, the results appearing in more than 5000 published papers. With the advent of computerized systems for analyzing two-dimensional (2-D) gel images and constructing spot databases, it is also possible to plan and assemble integrated bodies of information describing the appearance and regulation of thousands of protein gene products [5, 6]. Creating such databases involves amassing and organizing quantitative data from thousands of 2-D gels, and requires a substantial commitment in technology and resources.

Given the long-term effort required to develop a protein database, the choice of a biological system takes on considerable importance. While *in vitro* systems are ideal for answering many experimental questions, especially in cancer research and genetics, our experience with cell cultures and tissue samples suggests that some *in vivo* approaches could have major advantages. In particular, we have noticed that liver tissue samples from rats and mice appear to show greater quantitative reproducibility (in terms of individual protein expression) than replicate cell cultures. This is perhaps a natural result of the homeostasis maintained in a complete animal *vs.* the well-known variability of cell cultures, the latter due principally to differences in reagents (*e.g.*, fetal bovine serum), conditions (*e.g.*, pH) and genetic "evolution" of cell lines while in culture. It is also more difficult to generate adequate amounts of protein from cell culture systems (particularly with attached cells), forcing the investigator to resort to radioisotope-based or silver-based stain-detection methods. While these methods are more sensitive (sometimes much more sensitive) than the Coomassie Brilliant Blue (CBB) stain typically used for protein detection in "large" protein samples, they are generally more variable, more labor-intensive and, in the case of radiographic methods, may generate highly "noisy" images, due to the properties of the films used. By contrast, large protein samples can easily be prepared from liver using urea/Nonidet P-40 (NP-40) solubilization and stained with CBB, which has the advantage of being easily reproducible [8]. Finally, there remains the question of the "truthfulness" of many *in vitro* systems as compared to their *in vivo* analogs; how great are the changes caused by the introduction into a cul-

0173-0835/91/1111-0907 $3.50+.25/0

ture and the associated shift to strong selection for growth, and how do these affect experimental outcomes? Hence the apparent advantages of *in vitro* systems, in terms of experimental manipulation, may be counterbalanced by other factors relating to 2-D data quality.

There is a second important class of reasons for exploring the use of an *in vivo* biological system such as the liver. Historically, there have been two broad approaches to the mechanistic dissection of biochemical processes in intact cellular systems: genetics (a search for informative mutants) and the use of chemical agents (drugs and chemical toxins). Both approaches help us to understand complex systems by disrupting some specific functional element and showing us the result. With the development of techniques for genetic manipulation and cloning, the genetic approach can be effectively applied either *in vitro* or *in vivo*, although the *in vitro* route is usually quicker. The chemical approach can also be applied to either sort of biological system; here, however, the bulk of consistently acquired information is in experimental animals (rats and mice). While most biologists know a short list of compounds having specific, experimentally useful effects (*e.g.*, inhibitors of protein synthesis, ionophores, polymerase inhibitors, channel blockers, nucleotide analogs, and compounds affecting polymerization of cytoskeletal proteins), there is a much larger number of interesting chemically-induced effects, most of them characterized by toxicologists and pharmacologists in rodent systems. Just as a thorough genetic analysis would involve saturating a genome with mutations, it is possible to imagine a saturating number of drugs, the analysis of whose actions would reveal the complete biochemistry of the cell. While organized drug discovery efforts usually target specific desired effects, the nature of the process, with its dependence on screening large numbers of compounds, necessarily produces many unanticipated effects. It is therefore reasonable to suppose that the required broad range of compounds necessary to achieve "biochemical saturation" may be forthcoming; in fact, it may already exist among the hundreds of thousands of compounds that failed to qualify as drugs.

Among organs, the liver is an obvious choice for the study of chemical effects because of its well-known plasticity and responsiveness. The brain appears to be quite plastic (*e.g.* [7]), but it is a complicated mixture of cell types requiring skillful dissection for most experiments. The kidney, while quite responsive, also presents a potentially confounding mixture of cell types. The liver, by contrast, is made up of one predominant cell type which is easy to solubilize: the hepatocyte, representing more than 95% of its mass. Most importantly, the liver performs many homeostatic functions that require rapid modulation of gene expression. It appears that most chemical agents tested affect gene expression in the liver at some dosage (N. Leigh Anderson, unpublished observations), an interesting contrast to our earlier work with lymphocytes, for example, which seem to be much less responsive. Such results conform to the expectation that cells with a homeostatic, physiological role should be more plastic than cells differentiated for a purpose dependent on the action of a limited number of specific genes.

The liver also allows the parallels between *in vitro* and *in vivo* systems to be examined in detail. Significant progress has been made in the development of mouse, rat and human hepatocyte culture systems, as well as in precision-cut tissue slices. Using such an array of techniques, it is possible to assemble a matrix of mammalian systems including mouse and rat *in vivo* on one level and mouse, rat and human *in vitro* on a second level, and to compare effects between species and between systems. This approach allows us to draw informed conclusions regarding the biochemical "universality" of biological responses among the mammals and to offer some insight into the validity of *in vitro* approaches for toxicological screening. We believe this data will be necessary if *in vitro* alternatives are to achieve wide usage in government-mandated safety testing of drugs, consumer products and industrial and agricultural chemicals.

A number of interesting studies have been published using 2-D mapping to examine effects in the rodent liver. A number of investigarors have made use of the technique to screen for existing genetic variants [8–11] or induced mutations [12–14], mainly in the mouse. This work builds on the wealth of genetic information available on the mouse and its established position as a mammalian mutation-detection system. While some studies of chemical effects have been undertaken in the mouse [15–17], most have used the rat [18–23]. The examination of the cytochrome p-450 system, in particular, has been carried out almost exclusively on the rat [24, 25].

These considerations lead us to conclude that rodent liver offers the best opportunity to systematically examine an array of gene regulation systems, and ultimately to build a predictive model of large-scale mammalian gene control. The basic underlying foundation of such a project is a reliable, reproducible master 2-D pattern of liver, to which ongoing experimental results can be referred. In this paper, we report such a master pattern for the acidic and neutral proteins of rat liver (pattern F344MST3). In future, this master will be supplemented by maps of basic proteins, and analogous maps of mouse and human liver.

2 Materials and methods

2.1 Sample preparation

Liver is an ideal sample material for most biochemical studies, including 2-D analysis. A sample is taken of approximately 0.5 g of tissue from the apical end of the left lobe of the liver. Solubilization is effected as rapidly as practical; a delay of 5–15 min appears to cause no major alteration in liver protein composition if the liver pieces are kept cold (*e.g.*, on ice) in the interim. In the solubilization process, the liver sample is weighed, placed in a glass homogenizer (*e.g.*, 15 mL Wheaton); 8 volumes of solubilizing solution[*]

[*] The solubilizing solution is composed of 2% NP-40 (Sigma), 9 M urea (analytical grade, *e.g.*, BDH or Bio-Rad), 0.5% dithiothreitol (DTT: Sigma) and 2% carrier ampholytes (pH 9–11 LKB: these come as a 20% stock solution, so 2% final concentration is achieved by making the final solution 10% 9–11 Ampholine by volume). A large batch of solubilizer (several hundred mL) is made and stored frozen at −80°C in aliquots sufficient to provide enough for one day's estimated sample preparation requirement. The solution is never allowed to become warmer than room temperature at any stage during preparation or thawing for use, since heating of concentrated urea solutions can produce contaminants that covalently modify proteins, producing artifactual charge shifts. Once thawed, any unused solubilizer is discarded.

...ded (*i.e.*, 4 mL per 0.5 g tissue) and the mixture is ho...genized using first the loose- and then then the tight-fit...g glass pestle. This takes approximately 5 strokes with ...ch pestle and is carried out at room temperature because ...ea would crystallize out in the cold. Once the liver sample ...thoroughly homogenized in the solubilizer, it is assumed ...at all the proteins are denatured (by the chaotropic effect ...the urea and NP-40 detergent) and the enzymes inacti...ted by the high pH (~9.5). Therefore these samples may ...kept at room temperature until they can be centrifuged ...frozen as a group (within several hours of preparation). ...e samples are centrifuged for 6×10^6 *g* min (*e.g.*, 500 000 ...*g* for 12 min using a Beckman TL-100 centrifuge). The ...ntrifuge rotor is maintained at just below room tempera...re (*e.g.*, 15–20 °C), but not too cold, so as to prevent the ...ecipitation of urea. The centrifuge of choice is a Beckman ...L-100 because of the sample tube sizes available, but any ...ltracentrifuge accepting smallish tubes will suffice. When ...n appropriate centrifuge is not available near the site of ...ample preparation, samples can be frozen at –80 °C and ...hawed prior to centrifugation and collection of superna...nts. Each supernatant is carefully removed following cen...rifugation and aliquoted into at least 4 clean tubes for stor...ge. This is done by transferring all the supernatant to one ...lean tube, mixing this gently (to assure homogeneous ...omposition) and then dividing it into 4 aliquots. The ali...uots are frozen immediately at –80°C. These multiple ali...uots can provide insurance against a failed run or a freezer ...breakdown.

2.2 Two-dimensional electrophoresis

...ample proteins are resolved by 2-D electrophoresis using ...he 20 × 25 cm Iso-Dalt® 2-D gel system ([26–29]; pro...duced by LSB and by Hoefer Scientific Instruments, San ...Francisco) operating with 20 gels per batch. All first-dimen...sional isoelectric focusing (IEF) gels are prepared using the ...same single standardized batch of carrier ampholytes ...(BDH 4–8A in the present case, selected by LSB's batch-...testing program for rat and mouse database work**). A 10 ...µL sample of solubilized liver protein is applied to each gel, ...and the gels are run for 33 000 to 34 500 volt-hours using a ...progressively increasing voltage protocol implemented by ...a programmable high-voltage power supply. An Ange...lique™ computer-controlled gradient-casting system (pro...duced by LSB) is used to prepare second-dimensional sod...ium dodecyl sulfate (SDS) polyacrylamide gradient slab ...gels in which the top 5% of the gel is 11%T acrylamide, and ...the lower 95% of the gel varies linearly from 11% to 18%T.

...This system has recently been modified so as to employ a ...commercially available 30.8%T acrylamide/*N,N'*-methyle...nebisacrylamide prepared solution (thus avoiding the han...dling of the solid acrylamide monomer) and three addi...tional stock solutions: buffer (made from Sigma pre-set ...Tris), persulfate and *N,N,N',N'*-tetramethylethylenedi...amine (TEMED). Each gel is identified by a computer-...printed filter paper label polymerized into the lower left cor...ner of the gel. First-dimensional IEF tube gels are loaded directly (as extruded) onto the slab gels without equilibration, and held in place by polyester fabric wedges (Wedgies™, produced by LSB) to avoid the use of hot agarose. Second-dimensional slab gels are run overnight, in groups of 20, in cooled DALT tanks (10 °C) with buffer circulation. All run parameters, reagent source and lot information, and notations of deviation from expected results are entered by the technician responsible on a detailed, multi-page record of the experiment.

2.3 Staining

Following SDS-electrophoresis, slab gels are stained for protein using a colloidal Coomassie Blue G-250 procedure in covered plastic boxes, with 10 gels (totalling approximately 1 L of gel) per box. This procedure (based on the work of Neuhoff [30, 31]) involves fixation in 1.5 L of 50% ethanol and 2% phosphoric acid for 2 h, three 30 min washes, each in 2 L of cold tap water, and transfer to 1.5 L of 34% methanol, 17% ammonium sulfate and 2% phosphoric acid for 1 h, followed by the addition of a gram of powdered Coomassie Blue G-250 stain. Staining requires approximately 4 days to reach equilibrium intensity, whereupon gels are transferred to cool tap water and their surfaces rinsed to remove any particulate stain prior to scanning. Gels may be kept for several months in water with added sodium azide. The water washes remove ethanol that would dissolve the stain (and render the system noncolloidal, with high backgrounds). The concentrated ammonium sulfate and methanol solution is diluted by equilibration with the water volume of the gels to automatically achieve the correct final concentrations for colloidal staining. Practical advantages of this staining approach can be summarized as follows: (i) the low, flat background makes computer evaluation of small spots (max OD < 0.02) possible, especially when using laser densitometry; (ii) up to 1500 spots can be reliably detected on many gels (*e.g.*, rat liver) at loadings low enough to preserve excellent resolution; and (iii) reproducibility appears to be very good: at least several hundred spots have coefficients of reproducibility less than 15%. This value is at least as good as previous CBB methods, and significantly better than many silver stain systems.

2.4 Positional standardization

The carbamylated rabbit muscle creatine phosphokinase (CPK) standards [32] are purchased from Pharmacia and BDH. Amino acid compositions, and numbers of residues present in proteins used for internal standardization, are taken from the Protein Identification Resource (PIR) sequence database [33].

2.5 Computer analysis

Stained slab gels are digitized in red light at 134 micron resolution, using either a Molecular Dynamics laser scanner (with pixel sampling) or an Eikonix 78/99 CCD scanner. Raw digitized gel images are archived on high-density DAT tape (or equivalent storage media) and a greyscale videoprint prepared from the raw digital image as hard-copy backup of the gel image. Gels are processed using the Kepler® software system (produced by LSB), a commercially available workstation-based software package built on

**This material (succeeding certified batches of which are available from Hoefer Scientific Instruments) has the most linear pH gradient produced by any ampholyte tested except for the Pharmacia wide range (which has an unacceptable tendency to bind high-molecular weight acidic proteins, causing them to streak).

some of the principles of the earlier TYCHO system [34–41]. Procedure PROC008 is used to yield a spotlist giving position, shape and density information for each detected spot. This procedure makes use of digital filtering, mathematical morphology techniques and digital masking to remove the background, and uses full 2-D least-squares optimization to refine the parameters of a 2-D Gaussian shape for each spot. Processing parameters and file locations are stored in a relational database, while various log files detailing operation of the automatic analysis software are archived with the reduced data. The computed resolution and level of Gaussian convergence of each gel are inspected and archived for quality control purposes.

Experiment packages are constructed using the Kepler experiment definition database to assemble groups of 2-D patterns corresponding to the experimental groups (*e.g.*, treated and control animals). Each 2-D pattern is matched to the appropriate "master" 2-D pattern (pattern F344MST3 in the case of Fischer 344 rat liver), thereby providing linkage to the existing rodent protein 2-D databases. The software allows experiments containing hundreds of gels to be constructed and analyzed as a unit, with up to 100 gels displayed on the screen at one time for comparative purposes and multiple pages to accommodate experiments of > 1000 gels. For each treatment, proteins showing significant quantitative differences *vs.* appropriate controls are selected using group-wise statistical parameters (*e.g.*, Student's t-test, Kepler® procedure STUDENT). Proteins satisfying various quantitative criteria (such as $P < 0.001$ difference from appropriate controls) are represented as highlighted spots onscreen or on computer-plotted protein maps and stored as spot populations (*i.e.*, logical vectors) in a liver protein database. Quantitative data (spot parameters, statistical or other computed values) are stored as real-valued vectors in the database. Analysis of coregulation is performed using a Pierson product-moment correlation (Kepler procedure CORREL) to determine whether groups of proteins are coordinately regulated by any of the treatments. Such groups can be presented graphically on a protein map, and reported together with the statistical criteria used to assess the level of coregulation. Multivariate statistical analysis (*e.g.*, principal components' analysis) is performed on data exported to SAS (SAS Institute).

2.6 Graphical data output

Graphical results are prepared in GKS and translated within Kepler® into output for any of a variety of devices. Linedrawing output is typically prepared as Postscript and printed on an Apple Laserwriter. Detailed maps presented here have been generated using an ultra-high-resolution Postscript-compatible Linotronic output device. Greyscale graphics are reproduced from the workstation screen using a Seikosha videoprinter. Patterns are shown in the standard orientation, with high molecular mass at the top and acidic proteins to the left.

2.7 Experiment LSBC04

In the study described here 12-week-old Charles River male F344 rats were used. Diets were prepared at LSB, based on a Purina 5755M Basal Purified Diet. Lovastatin and cholestyramine were obtained as prescription pharmaceuticals, ground and mixed with the diet at concentrations of 0.075% and 1%, respectively. The high cholesterol diet was Purina 5801M-A (5% cholesterol plus 1% sodium cholate in the control diet). Animal work was carried out by Microbiological Associates (Bethesda, MD). Animals were acclimatized for one week on the control diet, fed test or control diets for one week, and sacrificed on day 8. Average daily doses of lovastatin and cholestyramine in appropriate groups were 37 mg/kg/day and 5 g/kg/day, respectively, based on the weight of the food consumed. Liver samples were collected and prepared for 2-D electrophoresis according to the standard liver protocol (homogenization in 8 volumes of 9 M urea, 2% NP-40, 0.5% dithiothreitol, 2% LKB pH 9–11 carrier ampholytes, followed by centrifugation for 30 min at 80000 × g). Kidney, brain and plasma samples were frozen. Gels were run as described above, and the data was analyzed using the Kepler® system. Gels were scaled, to remove the effect of differences in protein loading, by setting the summed abundances of a large number of matched spots equal for each gel (linear scaling).

3 Results and discussion

3.1 The rat liver protein 2-D map

F344MST3 is a standard 2-D pattern of rat liver proteins, based on the Fischer 344 strain. This pattern was initiated from a single 2-D gel and extensively edited in an experiment comparing it to a range of protein loads, so as to include both small spots and well-resolved representations of high-abundance spots. More than 700 rat liver 2-D patterns have been matched to F344MST3 in a series of drug effects and protein characterization experiments, and numerous new spots (induced by specific drugs, for instance) have been added as a result. A modified version including additional spots present in the Sprague-Dawley outbred rat has also been developed (data not shown). Figure 1 shows a greyscale representation and Fig. 2 a schematic plot of the master pattern. More than 1200 spots are included, most of which are visible on typical gels loaded with 10 µL of solubilized liver protein prepared by the standard method and stained with colloidal Coomassie Blue. Master spot numbers (MSN's) have been assigned to all proteins, and appear in the following figures, each showing one quadrant of the pattern. Figure 3 shows the upper left (acidic, high molecular mass) quadrant, Fig. 4 the upper right (basic, high molecular mass) quadrant, Fig. 5 the lower left (acidic, low molecular mass) quadrant, and Fig. 6 the lower right (basic, low molecular mass) quadrant. The quadrants overlap as an aid to moving between them. The gel position (in 100 micron units), isoelectric point (relative to the CPK internal p*I* standards) and SDS molecular mass (from the calibration curve in Fig. 8) are listed for each spot (Table 1). Because of the precision of the CPK-p*I* values, these parameters can be used to relate spot locations between gel systems more reliably than using p*I* measurements expressed as pH. A major objective of current studies is the identification of all major spots corresponding to known liver proteins, as well as rigorous definitions of subcellular organelle contents. Of particular interest to us is the parallel development of identifications in the rat and mouse liver maps, allowing detailed comparisons of gene expression effects in the two systems. The results of these studies will be presented systematically in a later edition of this database.

...we include here a useful series of 22 orienting identifications as an aid to other users of the rat liver pattern (Table ...).

...2 Carbamylated charge standards, computed pI's and molecular mass standardization

...e have previously shown that the use of a system of close-spaced internal pI markers (made by carbamylating a ...sic protein) offers an accurate and workable solution to ...e problem of assigning positions in the pI dimension [32]. ...he same system, based on 36 protein species made by carbamylating rabbit muscle CPK has been used here to as...ign pI's to most rat liver acidic and neutral proteins. The ...tandards were coelectrophoresed with total liver proteins, ...nd the standard spots added to a special version of the ...aster pattern F344MST3. The gel X-coordinates of all ...iver protein spots lying within the CPK charge train were ...hen transformed into CPK pI positions by interpolation ...etween the positions of immediately adjacent standards ...Table 1) using a Kepler[2] vector procedure.

...t has proven possible to compute fairly accurate pI values ...or many proteins from the amino acid composition [42]. We have attempted here to test a further elaboration of this ...pproach, in which we computed pI's for the CPK standards ...hemselves, based on our knowledge of the rabbit muscle CPK sequence and the fact that adjacent members of the ...harge train typically differ by blockage of one additional lysine residue (Table 3). We compared these values to similar computed pI's for an additional set of carbamylated standards made from human hemoglobin beta chains and a series of rat liver and human plasma proteins of known position and sequence (Fig. 7, Table 4). The result demonstrates good concordance between these systems. Two proteins show significant deviations: liver fatty-acid binding protein (FABP; #1 in Table 4) and protein disulphide isomerase (#20 in the table). The FABP spot present on F344MST3 may represent a charge-modified version of a more basic parent spot closer to the expected pI, not resolved in the IEF/SDS gel. Of particular importance is the fact that, by comparing computed pI's of sequenced but unlocated proteins with the CPK pI's, we can assign a probable gel location without making any assumptions regarding the actual gel pH gradient. This offers a useful shortcut, given the vagaries of pH measurement on small diameter IEF gels. We have used this approach to compute the CPK pI's of all rat and mouse proteins in the PIR sequence database, as an aid to protein identification (data not shown).

In order to standardize SDS molecular weight (SDS-MW), we have used a standard curve fitted to a series of identified proteins (Fig. 8). Rather than using molecular mass *per se*, we have elected to use the number of amino acids in the polypeptide chain, as perhaps a better indication of the length of the SDS-coated rod that is sieved by the second dimension slab. The resulting values were multiplied by 112 (the weighted average mass of amino acids in sequenced proteins) to give predicted molecular masses. Because we use gradient slabs, we have not constrained the fitted curve to conform to any predetermined model; rather we tried many equations and selected the best using the program "Tablecurve" on a PC. The equation chosen was $y = a + bx + c/x^2$, where y is the number of residues, x is the gel Y coordinate, a is 511.83, b is −0.2731 and c is 33183801. The resulting fit appears to be fairly good over a broad range of molecular mass.

3.3 An example of rat liver gene regulation: Cholesterol metabolism

Experiment LSBC04 was designed as a small-scale test of the regulation of cholesterol metabolism *in vivo* by three agents included in the diet: lovastatin (Mevacor[3], an inhibitor of HMG-CoA reductase); cholestyramine (a bile acid sequestrant that has the effect of removing cholesterol from the gut-liver recirculation); and cholesterol itself. The first two agents should lower available cholesterol and the third should raise it, allowing manipulation of relevant gene expression control systems in both directions. Such an experiment offers an interesting test of the 2-D mapping system since most of the pathway enzymes are present in low abundance, many are membrane-bound and difficult to solubilize, and the pathway itself is complex. Approximately 1000 proteins were separated and detected in liver homogenates. Twenty-one proteins were found to be affected by at least one treatment, and these could be divided into several coregulated groups.

3.3.1 MSN 413 (putative cytosolic HMG-CoA synthase) and sets of spots regulated coordinately or inversely

One group of spots (including a spot assigned to the cytosolic HMG-CoA synthase, MSN 413) showed the expected increase in abundance with lovastatin or cholestyramine, the synergistic further increase with lovastatin and cholestyramine, and a dramatic decrease with the high cholesterol diet. Spot number 413 is the most strongly regulated protein in the present experiment, showing a 5- to 10-fold induction after a 1 week treatment with 0.075% lovastatin and 1% cholestyramine in the diet (Figs. 9 and 10). Its expression follows precisely the expectation for an enzyme whose abundance is controlled by the cholesterol level; it is progressively increased from the control levels by cholestyramine, lovastatin and lovastatin plus cholestyramine, and it sinks below the threshold of detection in animals fed the high cholesterol diet. This spot has been tentatively identified as the cytosolic HMG-CoA synthase, based on a reaction with an antiserum to that protein provided by Dr. Michael Greenspan at Merck Sharp & Dohme Research Laboratories. This enzyme lies immediately before HMG-CoA reductase in the liver cholesterol biosynthesis pathway, and is known to be co-regulated with it. Spot 413 has an SDS molecular weight of about 54 000 and a CPK pI of −11.4, in reasonably close agreement with a molecular weight of 57300 and a CPK pI of −15.7 computed from the known sequence of the hamster enzyme [43].

Using a classical product-moment correlation test (Kepler procedure CORREL), a series of five additional spots was found to be coregulated with 413. The level of correlation was exceedingly high (> 95%). Two of these, 1250 and 933, are at similar molecular weights and approximately one charge more acidic than 413 (Fig. 9), indicating that they may be covalently modified forms of the 413 polypeptide. This suspicion is strengthened by the observation that both spots are also stained by the antibody to cytosolic HMG-CoA synthase. The remaining three correlated spots appear

to comprise an additional related pair (1253 and 1001) of around 40 kDa and a single spot (1119) of around 28 kDa. Because these two presumed proteins are present at substantially lower abundances than 413, and because the cytosolic HMG-CoA synthase is reported to consist of only one type of polypeptide, they are likely to represent other, very tightly coregulated enzymes. A second group of six spots was selected based on a regulatory pattern close to the inverse of that for spot 413 (MSN's 34, 79, 178, 182, 204, 347; data not shown). For these proteins, the lowest level of expression occurs with exposure to lovastatin plus cholestyramine and the highest level upon exposure to the high-cholesterol diet. Spots 182 and 79 are highly correlated and lie about one charge apart at the same molecular weight; they may thus be isoforms of a single protein. The other four spots probably represent additional enzymes or subunits.

3.3.2 MSN 235 and coregulated spots

A third group of five spots, mainly comprised of mitochondrial proteins including putative mitochondrial HMG-CoA synthase spots, showed a modest induction by lovastatin alone, but little or no effect with any of the other treatments (including the combination of lovastatin and cholestyramine; Fig. 12). This result is intriguing because lovastatin was expected to affect only the regulation of enzymes of cholesterol synthesis, which is entirely extra-mitochondrial. Three of the spots (235, 134, 144) form a closely-packed triad at approximately 30 kDa, and are likely to represent isoforms of one protein. All three spots are stained by an antibody to the mitochondrial form of HMG-CoA synthase obtained from Dr. Greenspan. Subcellular fractionation indicates a mitochondrial location. The other two spots (633 at about 38 kDa and 724 at about 69 kDa) are each present at lower abundance than the members of the triad.

3.3.3 An example of an anti-synergistic effect

A sixth spot (367) shows strong induction by lovastatin (two- to threefold), and about half as much induction with lovastatin plus cholestyramine, but without sharing the animal-animal heterogeneity pattern of the 235-set (Fig. 13). This protein is also mitochondrial, and represents the clearest example of an anti-synergistic effect of lovastatin and cholestyramine. The existence of such an effect demonstrates that lovastatin and cholestyramine do not act exclusively through the same regulatory pathway.

3.3.4 Complexity of the cholesterol synthesis pathway

Taken together, these results suggest that treatment with lovastatin alone can affect both cytosolic and mitochondrial pathways using HMG-CoA, while cholestyramine, on the other hand, either alone or in combination with lovastatin, produces a strong effect on the putative cytosolic pathway, but little or no effect on the putative mitochondrial pathway. An explanation for this difference may lie in lovastatin's effect on levels of HMG-CoA and related precursor compounds that are exchanged between the cytosol and the mitochondrion, whereas cholestyramine should affect only the cytosolic pathways directly controlled by cholesterol and bile acid levels. It remains to be explained why some proteins of the putative mitochondrial pathway are so much more variable in their expression in all groups. An examination of all the coregulated groups suggests that quantitative statistical techniques can extract a wealth of interesting information from large sets of reproducible gels. The abundance of spots in the 413 coregulation group, for example, shows an amazing level of concordance in their relative expression among the five individuals of the lovastatin and cholestyramine treatment group. This effect is not due to differences in total protein loading, since they have already been removed by scaling, and since proteins with quite different regulation patterns can be demonstrated (*e.g.*, Fig. 13). Such effects raise the possibility that many gene coregulation sets may be revealed through the study of a sufficiently large population of control animals (*i.e.*, without any experimental manipulation). This approach, exploiting natural biological variation in protein expression instead of drug effects, offers an important incentive for the construction of a large library of control animal patterns.

4 Conclusions

Because of the widespread use of rat liver in both basic biochemistry and in toxicology, there is a long-term need for a comprehensive database of liver proteins. The rat liver master pattern presented here has proven to be an accurate representation of this system, having been matched to more than 700 gels to date. As the number of proteins identified and the number of compounds tested for gene expression effects grows, we expect this database to contribute valuable insights into gene regulation. Its practical utility in several areas of mechanistic toxicology is already being demonstrated.

Received September 11, 1991

5 References

[1] O'Farrell, P., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[2] Klose, J., *Humangenetik* 1975, *26*, 231–243.

[3] Scheele, G. A., *J. Biol. Chem.* 1975, *250*, 5375–5385.

[4] Iborra, G. and Buhler, J. M., *Anal. Biochem.* 1976, *74*, 503–511.

[5] Anderson, N. G. and Anderson, N. L., *Behring. Inst. Mitt.* 1979, *63*, 169–210.

[6] Anderson, N. G. and Anderson, N. L., *Clin. Chem.* 1982, *28*, 739–748.

[7] Heydorn, W. E., Creed, G. J. and Jacobowitz, D. M., *J. Pharmacol. Exp. Therap.* 1984, *229*, 622–628.

[8] Anderson, N. L., Nance, S. L., Tollaksen, S. L., Giere, F. A. and Anderson, N. G., *Electrophoresis* 1985, *6*, 592–599.

[9] Racine, R. R. and Langley, C. H., *Biochem. Genet.* 1980, *18*, 185–197.

[10] Klose, J., *Mol. Evol.* 1982, *18*, 315–328.

[11] Neel, J. V., Baier, L., Hanash, S. and Erickson, R. P., *J. Hered.* 1985, *76*, 314–320.

[12] Marshall, R. R., Raj, A. S., Grant, F. J. and Heddle, J. A., *Can. J. Genet. Cytol.* 1983, *25*, 457–466.

[13] Taylor, J., Anderson, N. L., Anderson, N. G., Gemmell, A., Giometti, C. S., Nance, S. L. and Tollaksen, S. L., in: Dunn, M. J. (Ed.), *Electrophoresis '86*, Verlag Chemie, Weinheim 1986, pp. 583–587.

[14] Giometti, C. S., Gemmell, M. A., Nance, S. L., Tollaksen, S. L. and Taylor, J., *J. Biol. Chem.* 1987, *262*, 12764–12767.

[15] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., in: Galteau, M.-M. and Siest, G. (Eds.), *Progrès Récents en Electrophorèse Bidimensionelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253–260.

[16] Anderson, N. L., Swanson, M., Giere, F. A., Tollaksen, S., Gemmell, A., Nance, S. L. and Anderson, N. G., *Electrophoresis* 1986, *7*, 44–48.

Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L. and Anderson, N. G., *Fundam. Appl. Toxicol.* 1987, *8*, 39–50.

Anderson, N. L., in: *New Horizons in Toxicology*, Eli Lilly Symposium, 1991, in press.

Antoine, B., Rahimi-Pour, A., Siest, G., Magdalou, J. and Galteau, M. M., *Cell. Biochem. Funct.* 1987, *5*, 217–231.

Elliott, B. M., Ramasamy, R., Stonard, M. D. and Spragg, S. P. *Biochim. Biophys. Acta* 1986, *876*, 135–140.

Huber, B. E., Heilman, C. A., Wirth, P. J., Miller, M. J. and Thorgeirsson, S. S., *Hepatology* 1986, *6*, 209–219.

Wirth, P. J. and Vesterberg, O., *Electrophoresis* 1988, *9*, 47–53.

Witzmann, F. A. and Parker, D. N., *Toxicol. Lett.* 1991, *57*, 29–36.

Rampersaud, A., Waxman, D. J., Ryan, D. E., Levin, W. and Walz, F. G., Jr., *Arch. Biochem. Biophys.* 1985, *243*, 174–183.

Vlasuk, G. P. and Walz, F. G., Jr., *Anal. Biochem.* 1980, *105*, 112–120.

Anderson, N. G. and Anderson, N. L., *Anal. Biochem.* 1978, *85*, 331–340.

Anderson, N. L. and Anderson, N. G., *Anal. Biochem.* 1978, *85*, 341–354.

Anderson, L., Hofmann, J.-P., Anderson, E., Walker, B. and Anderson, N. G., in: Endler, A. T. and Hanash, S. (Eds.), *Two-Dimensional Electrophoresis*, VCH Verlagsgesellschaft, Weinheim 1989, pp. 288–297.

Anderson, L., *Two-Dimensional Electrophoresis: Operation of the ISO-DALT® System*. Large Scale Biology Press, Washington, DC 1988, ISBN 0-945532-00-8, 170pp.

Neuhoff, V., Stamm, R. and Eibl, H., *Electrophoresis* 1985, *6*, 427–448.

[31] Neuhoff, V., Arold, N., Taube, D. and Ehrhardt, W., *Electrophoresis* 1988, *9*, 255–262.

[32] Anderson, N. L. and Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.

[33] Sidman, K. E., George, D. E., Barker, W. C. and Hunt, L. T., *Nucl. Acids Res.* 1988, *16*, 1869–1871.

[34] Taylor, J., Anderson, N. L., Coulter, B. P., Scandora, A. E. and Anderson, N. G., in: Radola, B. J. (Ed.), *Electrophoresis '79*, de Gruyter, Berlin 1980, pp. 329–339.

[35] Taylor, J., Anderson, N. L. and Anderson, N. G., in: Allen, R. C. and Arnaud, P. (Eds.), *Electrophoresis '81*, de Gruyter, Berlin 1981, pp. 383–400.

[36] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. and Anderson, N. G., *Clin. Chem.* 1981, *27*, 1807–1820.

[37] Taylor, J., Anderson, N. L., Scandora, A. E., Jr., Willard, K. E. and Anderson, N. G., *Clin. Chem.* 1982, *28*, 861–866.

[38] Taylor, J., Anderson, N. L. and Anderson, N. G., *Electrophoresis* 1983, *4*, 338–345.

[39] Anderson, N. L. and Taylor, J., in: *Proceedings of the Fourth Annual Conference and Exposition of the National Computer Graphics Association*, Chicago, June 26–30, 1983, pp. 69–76.

[40] Anderson, N. L., Hofmann, J.-P., Gemmell, A. and Taylor, J., *Clin. Chem.* 1984, *30*, 2031–2036.

[41] Anderson, L., in: Schafer-Nielsen, C. (Ed.), *Electrophoresis '88*, VCH Verlagsgesellschaft, Weinheim 1988, pp. 313–321.

[42] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E. and Phillips, T. A., *Electrophoresis* 1989, *10*, 116–121.

[43] Gil, G., Goldstein, J. L., Slaughter, C. A. and Brown, M. S., *J. Biol. Chem.* 1986, *261*, 3710–3716.

6 Addendum 1: Figures 1–13

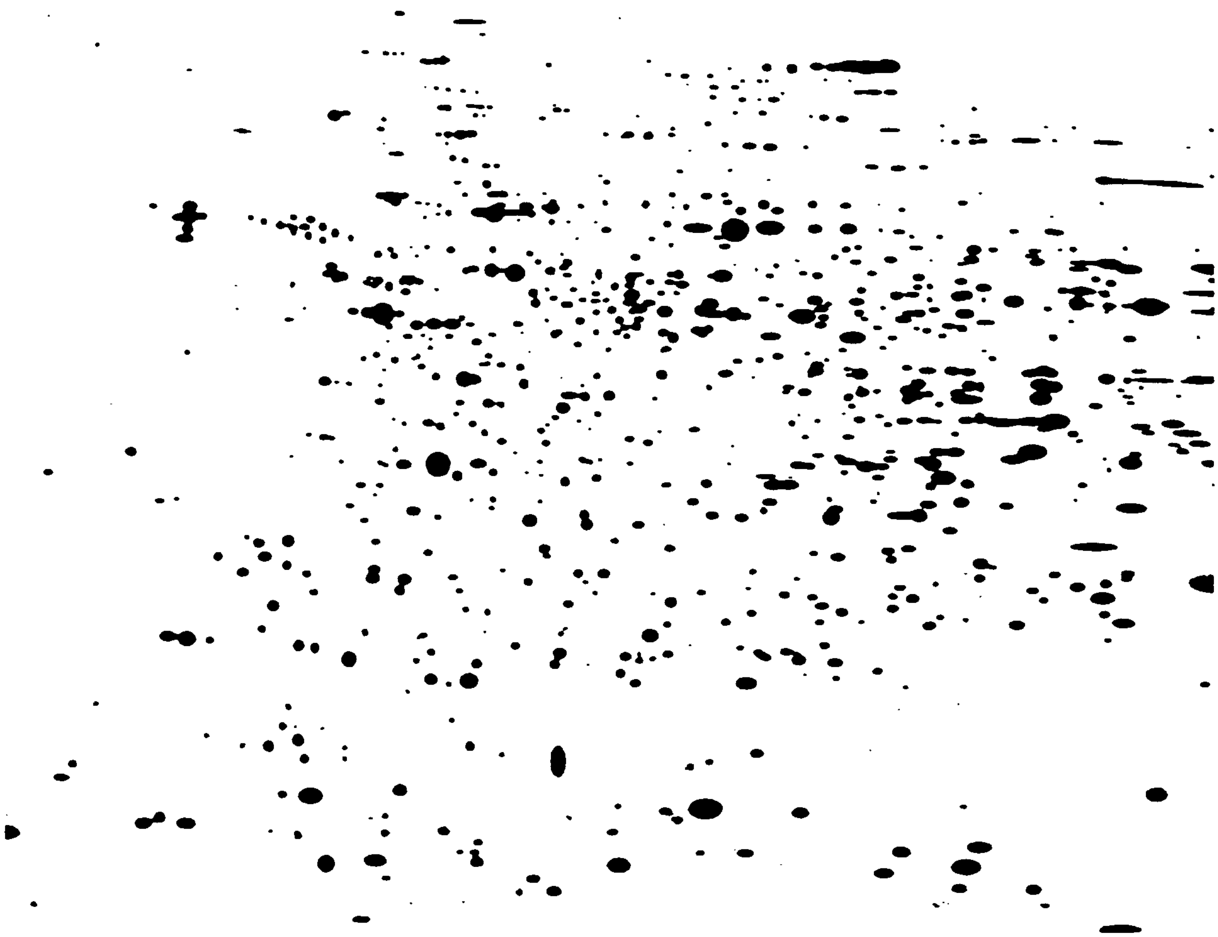
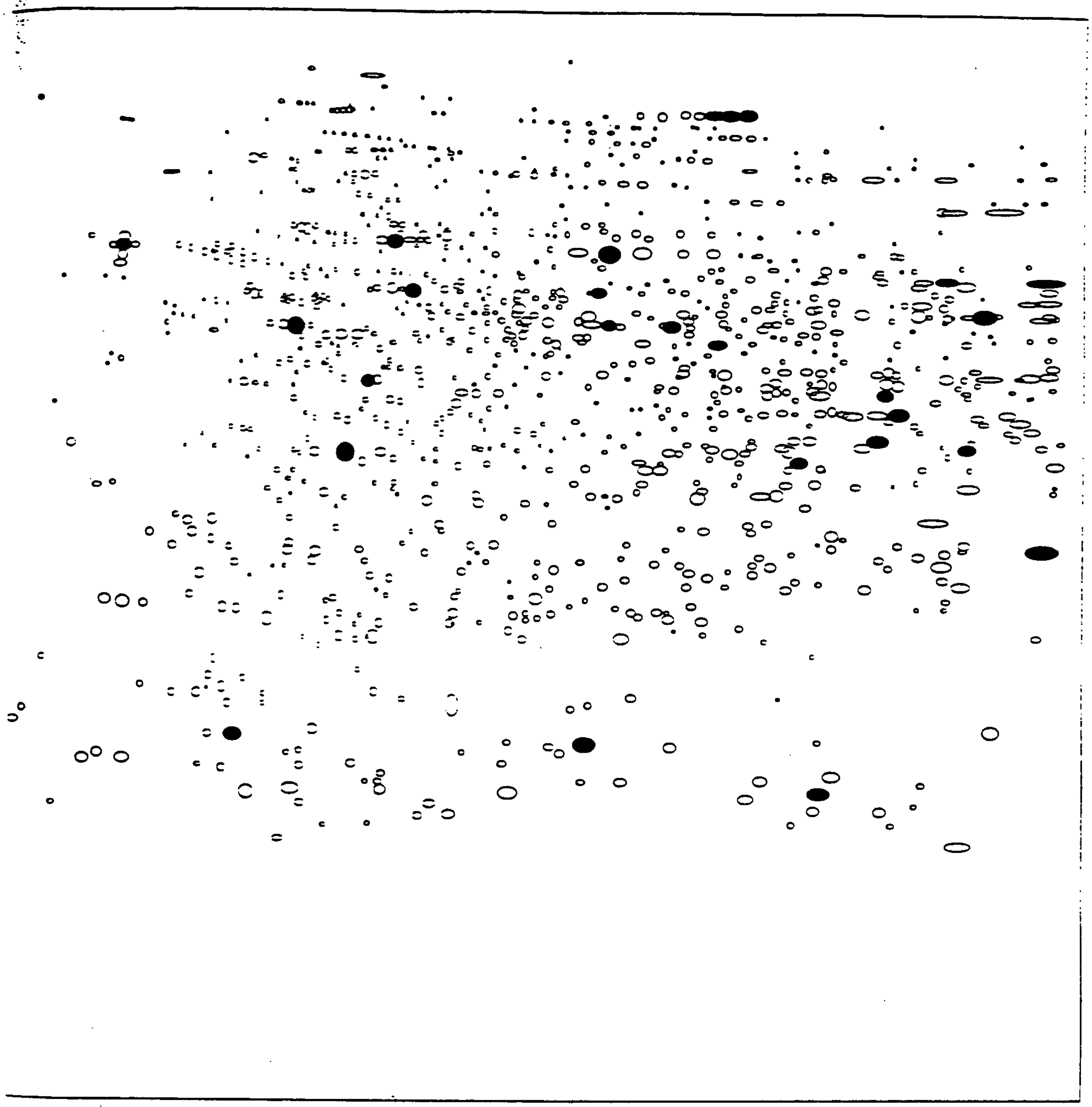


Figure 1. Synthetic representation of the standard rat liver 2-D master pattern, rendered as a greyscale image using a videoprinter.



...re 2. Schematic representation of the master pattern (the same as Fig. 1), useful as an aid in relating specific areas of Fig. 1 and the following detailed ...rants.

1

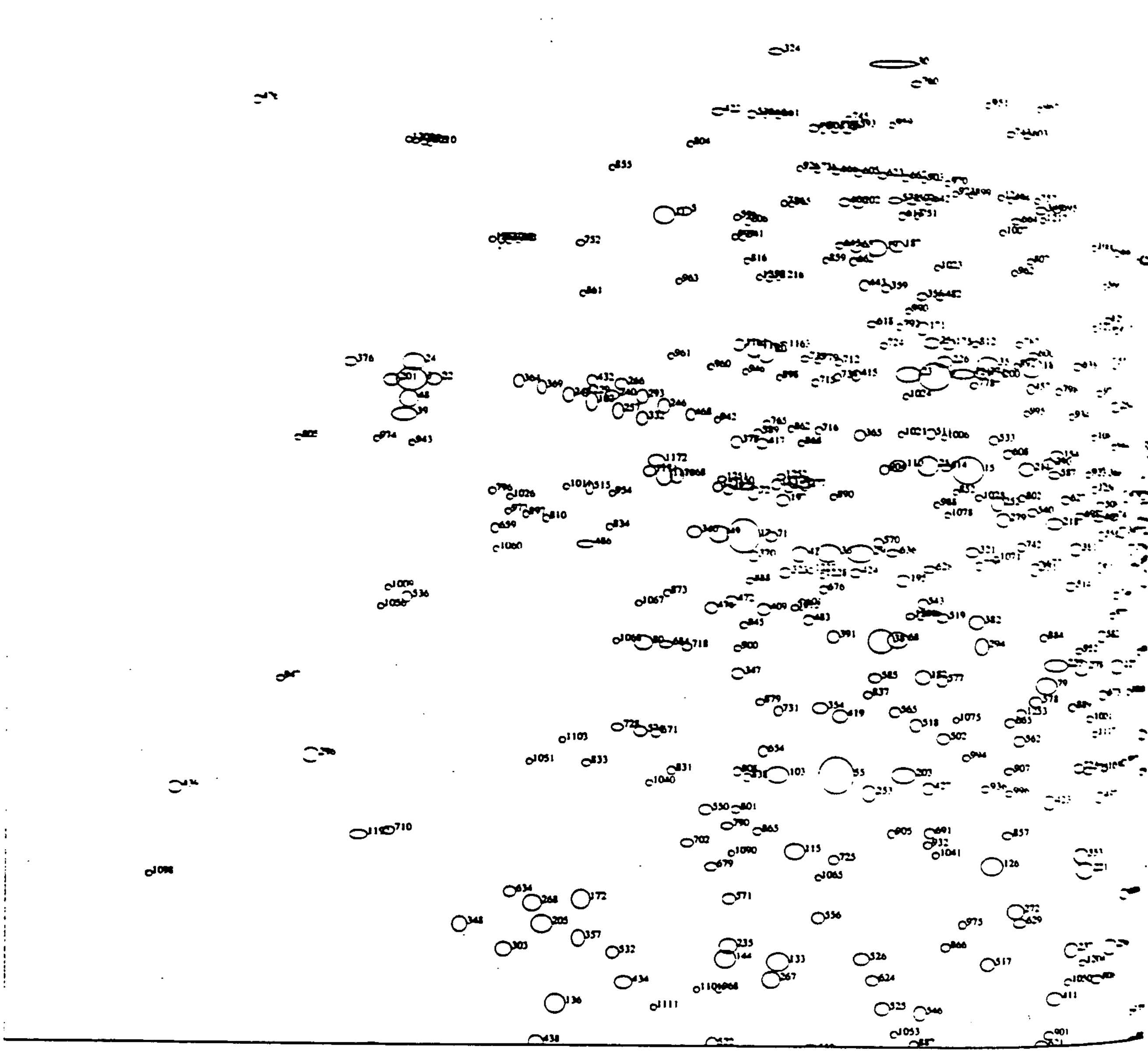
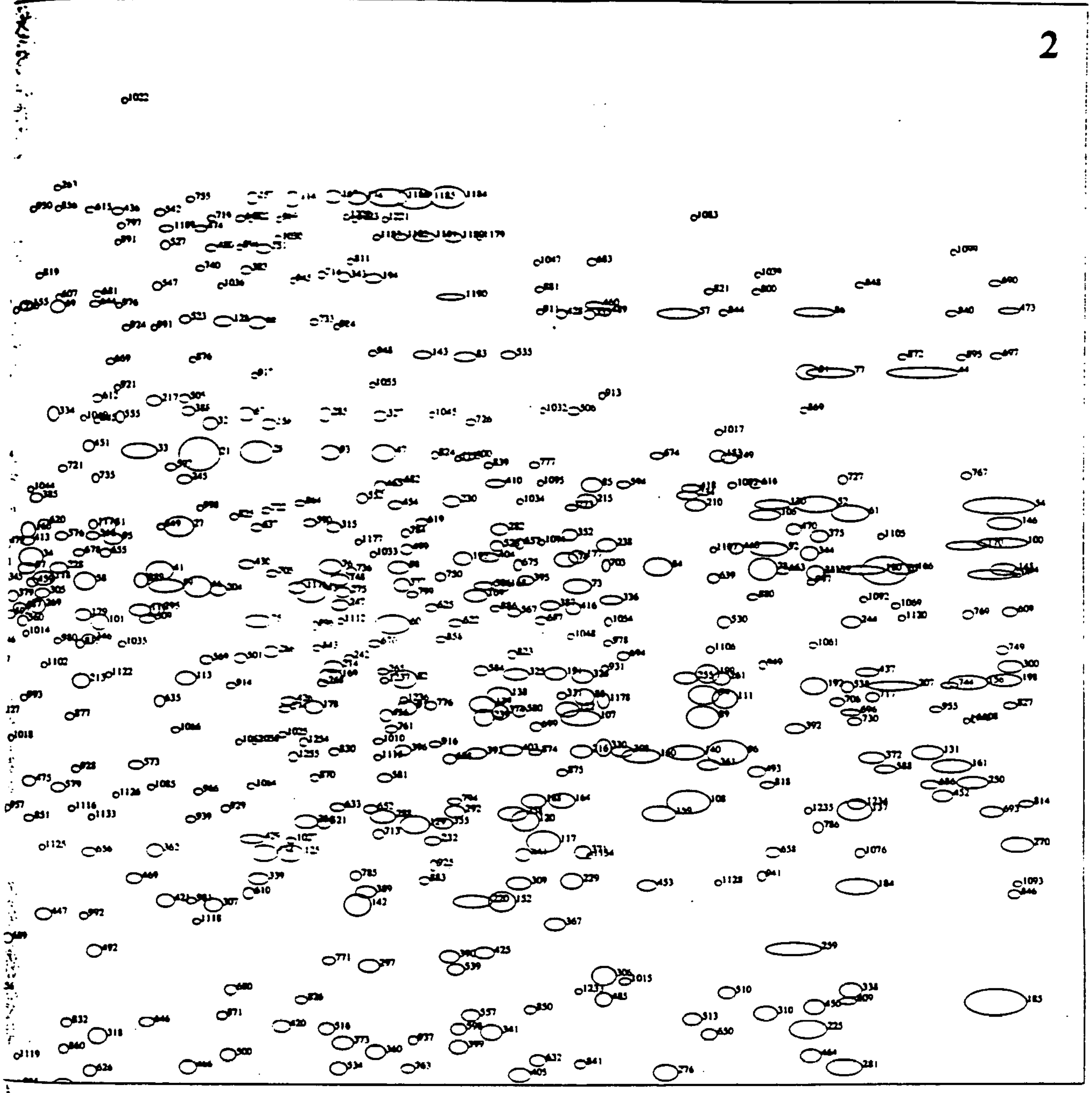


Figure 3. Upper left (high molecular weight, acidic) quadrant (#1) of the rat liver map, showing spot numbers.



gure 4. Upper right (high molecular weight, basic) quadrant (#2) of the rat liver map, showing spot numbers.

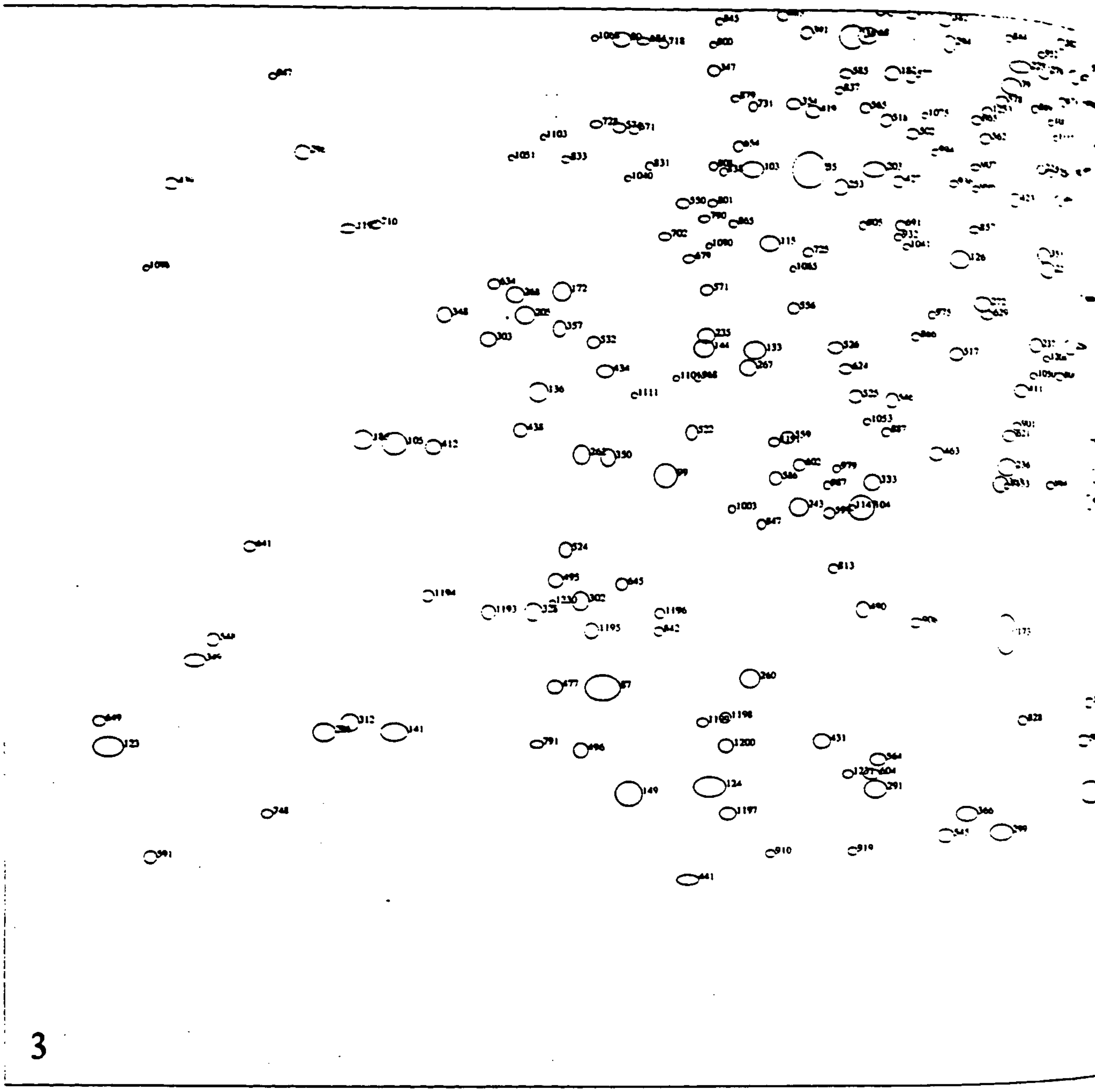
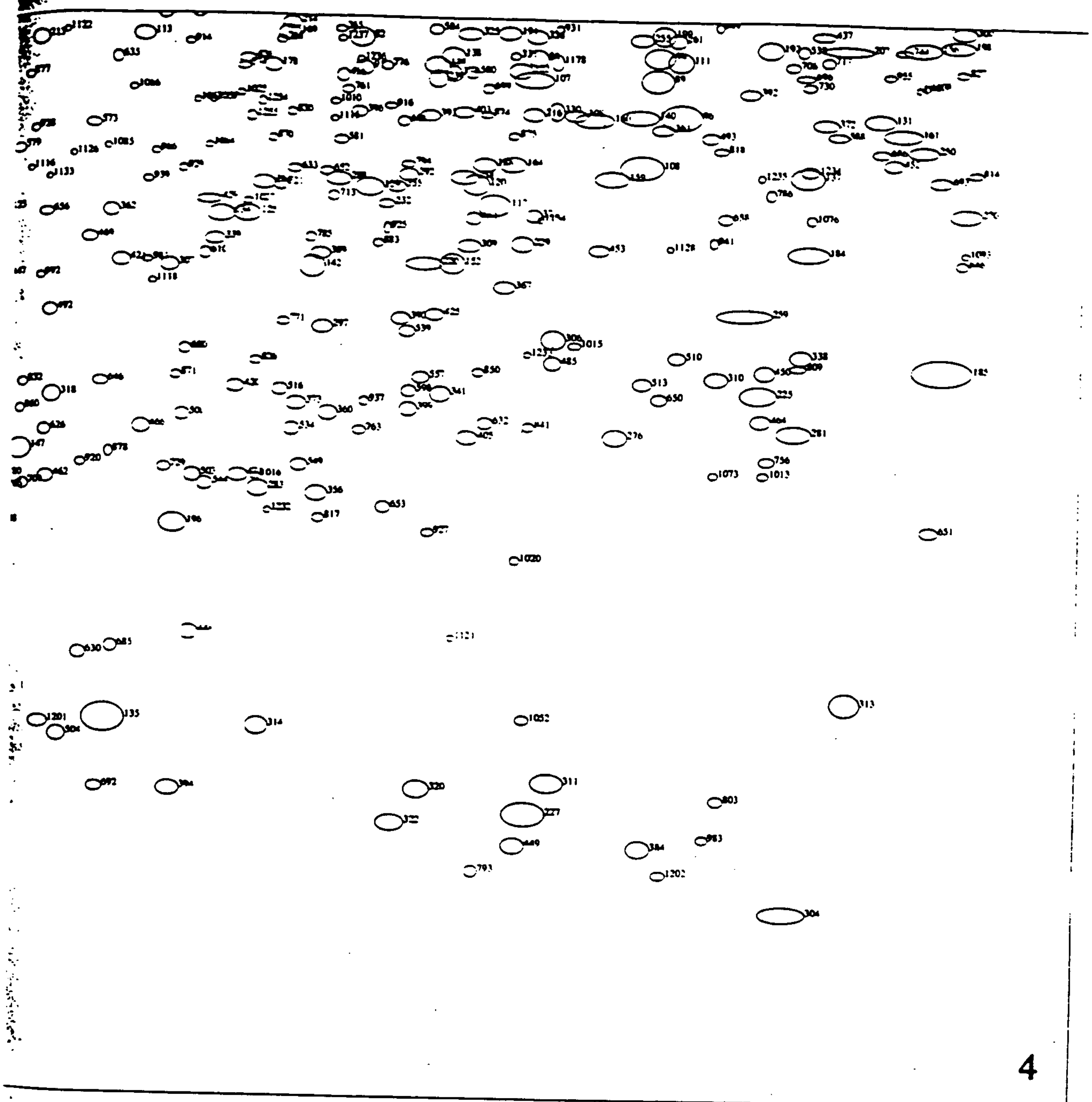
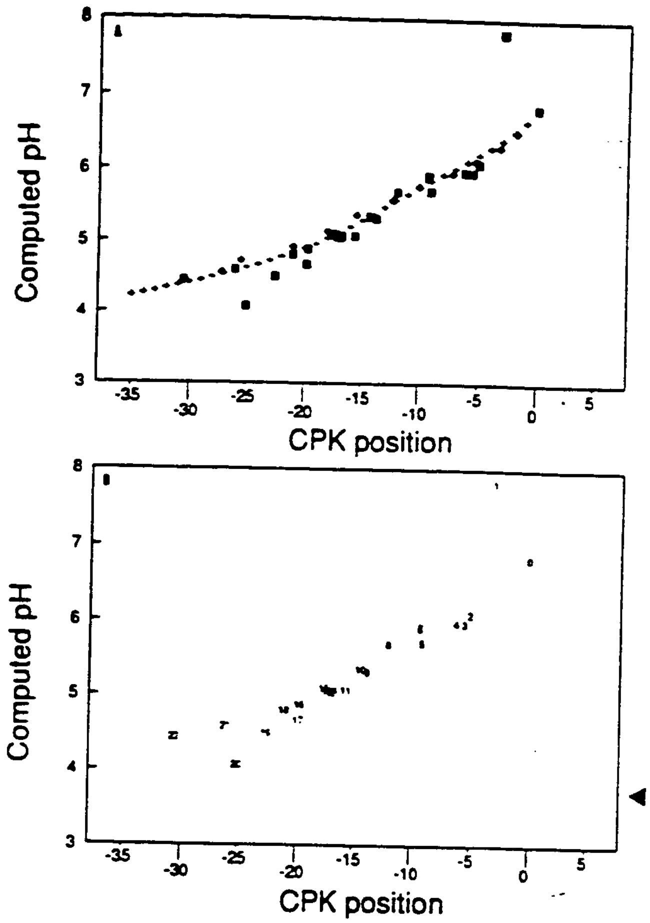


Figure 5. Lower left (low molecular weight, acidic) quadrant (#3) of the rat liver map, showing spot numbers.



...ure 6. Lower right (low molecular weight, basic) quadrant (#4) of the rat liver map, showing spot numbers.



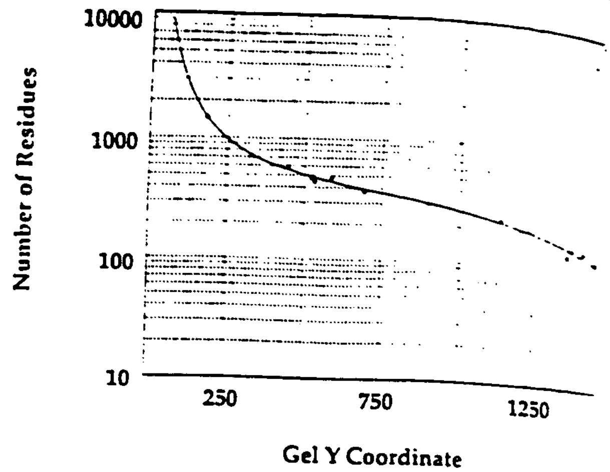


Figure 8. Plot of number of amino acids versus gel Y-position, with fitted curve used to predict molecular mass of unidentified proteins.

Figure 7. (a) Plot of computed isoelectric point versus gel X-position for two sets of carbamylated standard proteins (rabbit muscle CPK [+] and human hemoglobin β chain, filled diamonds) and several other proteins (shaded squares). (b) The identities of the various proteins represented by the squares are indicated by the numbers in corresponding positions on (a); these refer to Table 4.

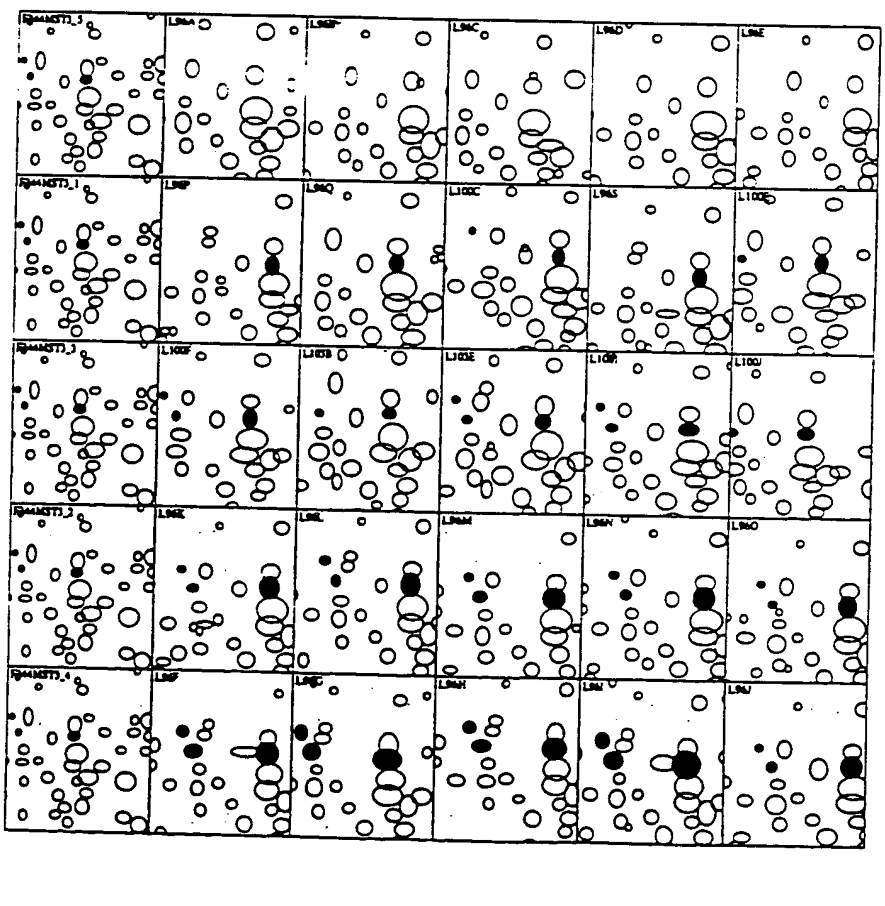


Figure 9. Montage showing effects in the region of MSN:413. The montage shows a small window into one portion of the 2-D pattern, one row of windows for each experimental group, and one panel for each gel in the experiment. The left-most pattern in each row is a group-specific copy of the master pattern followed by the patterns for the five individual rats in the group. The highlighted protein spots (filled circles) are spot 413 (on the right of each panel; identified as cytosolic HMG-CoA synthase) and two modified forms of it (1250 and 933). From the top, the rows (experimental groups) are: high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine.

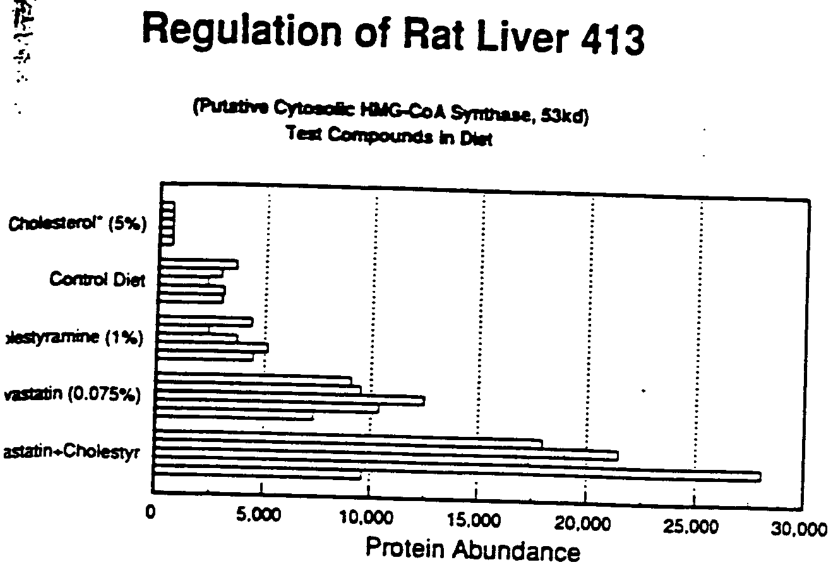


Figure 10. Bargraph showing the quantitative effects of various treatments on the abundance of MSN:413 (cytosolic HMG-CoA synthase) in the gels of Fig. 9.

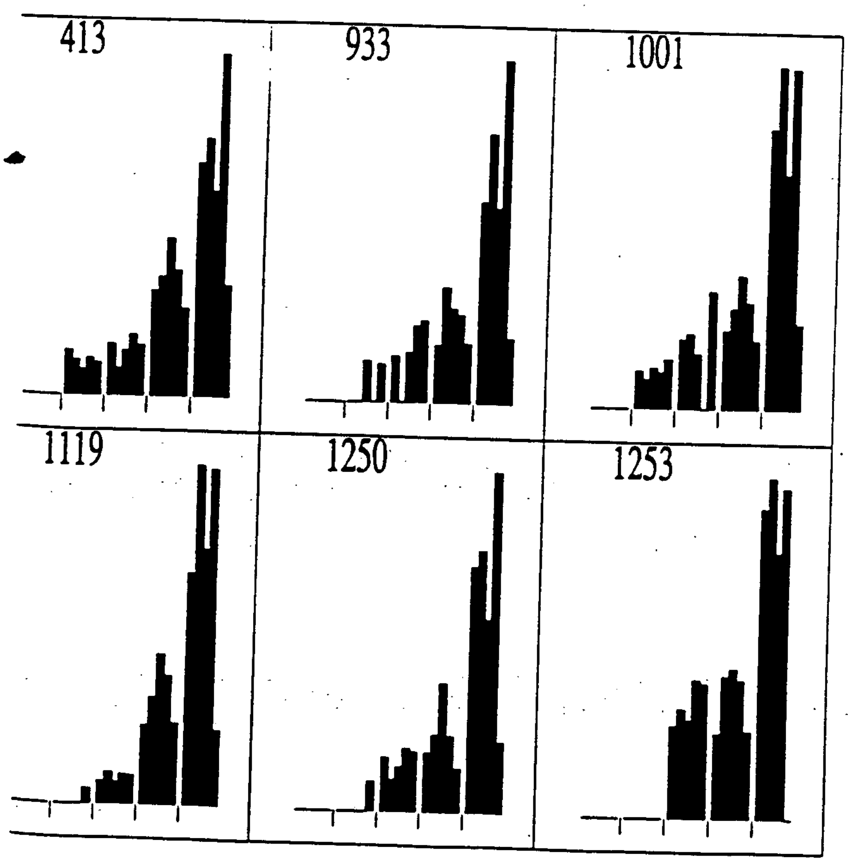


Figure 11. Bargraphs of a series of six coregulated spots including MSN:413. In the bargraphs, the abundances of the appropriate spot (master spot number shown at the top of the panel) in each animal are shown. The five five-animal groups are in the order (left to right): high cholesterol, controls, cholestyramine, lovastatin, and lovastatin plus cholestyramine. Each bar within a group represents one experimental animal liver (one 2-D gel). Note the correlated expression of the 6 spots, especially in the two far right (most strongly induced) groups.

Electrophoresis 1991, 12, 907–934

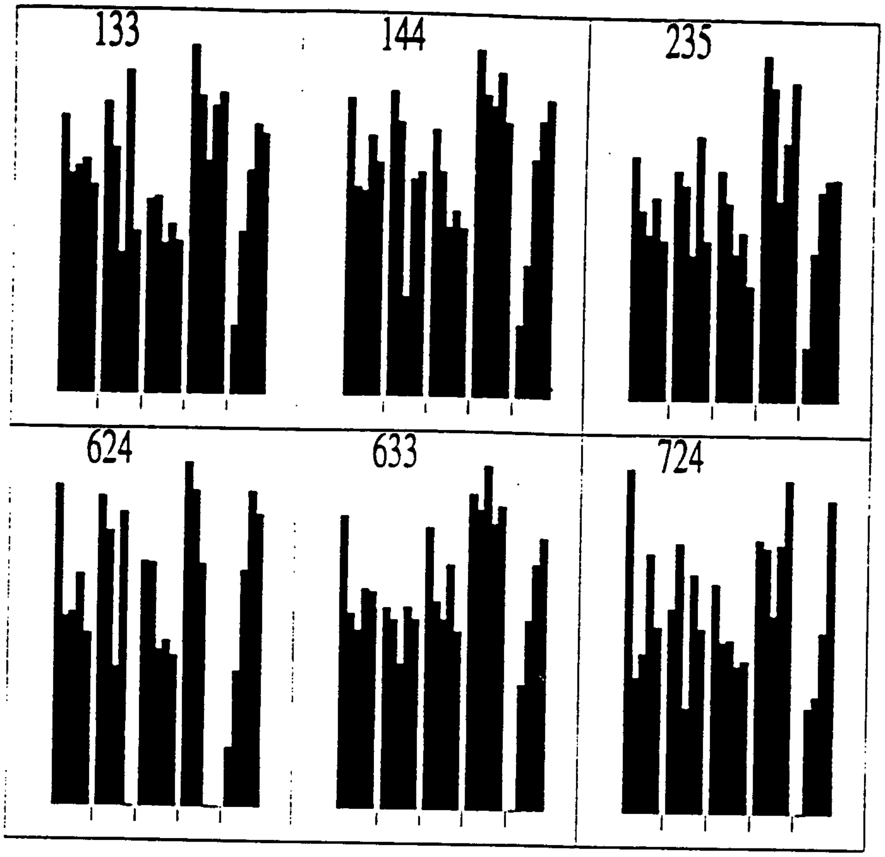


Figure 12. Data on a second coregulated group of spots, presented as in Fig. 11. The fourth experimental group (lovastatin) shows a modest induction, while the fifth group (lovastatin plus cholestyramine) does not.

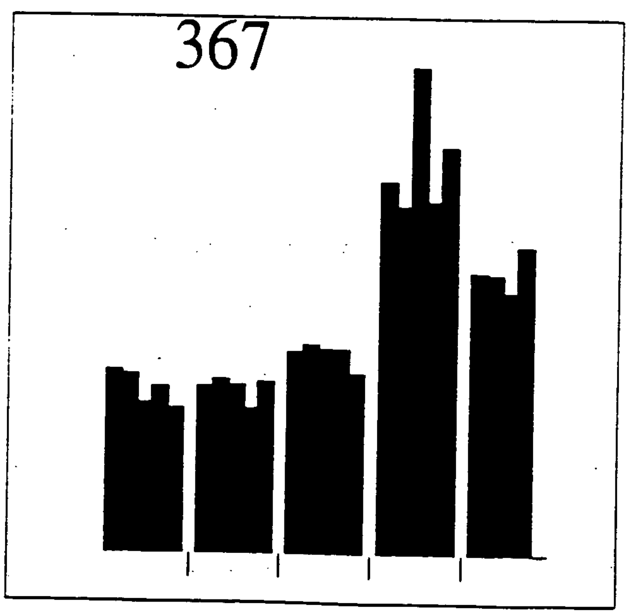


Figure 13. Data on spot MSN:367, presented as in Fig. 11. This protein shows unambiguously the anti-synergistic effect of lovastatin and cholestyramine (fifth group) as compared to lovastatin (fourth group). This response contrasts strongly with the regulation pattern seen in Fig. 11.

Addendum 2: Tables 1–4

Table 1. Master table of proteins in the rat liver database[a)]

MSN	X	Y	CPKpl	SDSMW
3	311	434	<-35.0	63,800
5	568	263	-24.3	102,900
8	812	426	-16.0	64,800
11	549	268	-25.2	101,000
15	845	520	-15.3	55,200
17	629	589	-21.6	50,000
18	906	414	-14.0	66,300
19	755	298	-17.5	90,200
20	649	403	-20.9	67,900
21	1204	448	-8.7	62,100
22	332	434	<-35.0	63,800
23	787	424	-16.6	65,000
24	313	417	<-35.0	66,000
25	807	516	-16.1	55,500
27	1184	524	-9.0	54,900
28	1263	446	-8.0	62,400
29	743	605	-17.8	49,000
30	768	112	-17.2	348,600
32	1216	417	-8.6	66,000
33	1145	445	-9.5	62,500
34	1037	555	-11.3	52,400
35	863	412	-14.9	66,600
36	712	606	-18.7	48,900
38	763	694	-17.3	43,800
39	304	470	<-35.0	59,800
41	1165	569	-9.2	51,400
42	684	607	-19.6	48,800
43	1318	589	-7.3	50,000
44	1924	362	-0.1	74,600
46	1203	586	-8.7	50,200
47	1391	447	-6.3	62,300
48	309	454	<-35.0	61,500
49	605	587	-22.5	50,100
50	621	535	-21.8	53,900
51	1113	522	-10.0	55,000
52	1820	499	-0.9	57,000
53	725	177	-18.3	170,800
54	2001	500	>0.0	56,900
55	722	830	-18.4	37,300
56	678	533	-19.8	54,100
57	1682	302	-2.5	89,000
58	1091	580	-10.3	50,600
59	1171	585	-9.2	50,300
60	1400	624	-6.2	47,800
61	1853	508	-0.6	56,200
62	1888	567	-0.4	51,500
65	735	297	-18.1	90,500
66	1263	312	-8.0	85,900
67	1252	407	-8.1	67,300
68	779	692	-16.8	43,900
69	1064	296	-10.8	90,800
71	656	589	-20.6	50,000
72	638	545	-21.2	53,100
73	1582	583	-3.6	50,400
74	1570	556	-3.8	52,300
75	1264	621	-8.0	48,000
76	1338	564	-7.0	51,800
77	1833	363	-0.8	74,400
78	1767	565	-1.5	51,700
79	925	738	-13.6	41,600
80	534	698	-26.1	43,600
81	1811	363	-1.0	74,500
82	1412	681	-6.0	44,500
83	1471	347	-5.0	77,500
84	1662	563	-2.7	51,800
85	1596	479	-3.4	58,900
86	1817	301	-0.9	89,100
87	516	1371	-27.0	17,400
88	1589	698	-3.5	43,600
89	1706	719	-2.2	42,500
90	651	329	-20.8	81,700
91	1415	710	-6.0	43,000
92	1773	545	-1.4	53,200
93	1338	446	-7.0	62,300
94	1708	696	-2.2	43,700
95	1119	536	-9.9	53,800
96	1731	756	-2.0	40,700
97	1033	566	-11.4	51,600
98	1406	565	-6.1	51,700
99	578	1149	-23.8	25,000
100	2004	538	>0.0	53,700
101	1106	623	-10.1	47,900
102	482	455	-28.5	61,300
103	665	830	-20.2	37,300
104	773	1182	-17.0	23,800
105	312	1117	<-35.0	26,100
106	1769	509	-1.5	56,100
107	1585	720	-3.6	42,500
108	1692	807	-2.4	38,300
109	1482	593	-4.8	49,700
110	778	516	-16.9	55,500
111	1728	700	-2.0	43,500
113	1191	680	-8.9	44,500
114	1298	185	-7.5	160,800
115	682	907	-19.6	34,100
116	1146	610	-9.5	48,700
117	1548	849	-4.1	36,500
118	1050	577	-11.1	50,800
120	1530	828	-4.3	37,400
121	838	423	-15.4	65,200
122	1572	712	-3.8	42,900
123	23	1433	<-35.0	15,300
124	621	1474	-21.9	13,900
125	1298	862	-7.5	36,000
126	872	921	-14.7	33,500
127	1000	717	-12.0	42,600
128	1229	311	-8.4	86,100
129	1422	832	-5.8	37,300
130	1776	499	-1.4	57,000
131	1930	757	-0.1	40,700
132	660	537	-20.4	53,800
133	666	1019	-20.2	29,700
134	1271	862	-7.9	36,000
135	1161	1389	-9.3	16,800
136	453	1063	-29.7	28,100
137	1858	823	-0.6	37,700
138	1504	697	-4.6	43,700
139	1488	707	-4.8	43,200
140	1689	756	-2.4	40,700
141	311	1417	<-35.0	15,800
142	1366	915	-6.7	33,800
143	1429	346	-5.7	77,900
144	615	1017	-22.1	29,800
145	2006	566	>0.0	51,600
146	2006	518	>0.0	55,300
147	1070	1108	-10.7	26,500
148	1347	578	-6.9	50,800
149	541	1481	-25.7	13,700
150	1645	760	-2.8	40,500
151	1269	236	-7.9	117,000
152	1507	911	-4.5	33,900
153	1722	448	-2.1	62,100
154	932	503	-13.5	56,600
155	1031	294	-11.4	91,400
156	1970	684	>0.0	44,400
157	1258	183	-8.1	162,400
158	1275	417	-7.8	65,900
159	1663	820	-2.6	37,800
160	1034	527	-11.4	54,600
161	1953	771	>0.0	40,000
162	1020	1482	-11.6	13,700
164	1566	806	-3.8	38,400
166	1905	565	-0.2	51,700
167	1340	181	-7.0	164,900
168	1506	583	-4.6	50,400
169	1338	678	-7.0	44,700
170	1969	541	>0.0	53,500
171	800	378	-16.3	71,800
172	476	958	-28.7	32,100
173	919	1314	-13.7	19,300
174	1364	183	-6.7	162,900
175	825	393	-15.7	69,300
177	1582	553	-3.6	52,600
178	1321	710	-7.2	43,000
179	1089	615	-10.4	48,300
180	1866	567	-0.5	51,600
181	411	295	-32.1	91,200
182	804	730	-16.2	42,000
184	1860	896	-0.6	34,500
185	1997	1017	>0.0	29,800
186	279	1113	<-35.0	26,300
187	773	296	-17.0	90,800
188	1538	807	-4.2	38,400
191	1560	674	-3.9	44,900
192	1818	687	-0.9	44,200
193	1469	555	-5.0	52,400
194	1380	266	-6.4	101,600
195	784	632	-16.7	47,300
196	1227	1185	-8.4	23,700
197	667	553	-20.1	52,600
198	2006	681	>0.0	44,500
199	1711	674	-2.2	44,900
200	872	424	-14.7	65,000
201	292	435	<-35.0	63,700
202	736	253	-18.0	107,800
203	786	829	-16.7	37,400
204	1224	589	-8.5	50,000
205	439	983	-30.9	31,100
206	1994	571	>0.0	51,300
207	1895	687	-0.3	44,200
208	240	1418	<-35.0	15,800
210	1700	499	-2.3	57,000
211	902	517	-14.1	55,400
213	1087	684	-10.4	44,400
214	1340	668	-7.0	45,200
215	1591	495	-3.5	57,300
216	1585	755	-3.6	40,700
217	1159	393	-9.3	69,300
218	931	572	-13.5	51,200
219	713	177	-18.7	170,500
220	1479	911	-4.9	33,900
221	965	927	-12.8	33,300
223	934	716	-13.5	42,700
225	1812	1045	-1.0	28,800
226	821	411	-15.8	66,800
227	1586	1483	-3.6	13,600
228	1065	567	-10.8	51,600
229	1577	890	-3.7	34,800
230	1458	496	-5.2	57,300
232	1440	849	-5.5	36,500
234	1692	489	-2.4	57,900
235	618	1004	-22.0	30,300
236	920	1138	-13.7	25,400
237	952	1008	-13.1	30,200
238	1611	541	-3.2	53,500
239	1489	720	-4.8	42,500
240	501	448	-27.7	62,100
241	1820	569	-0.9	51,400
242	1357	658	-6.8	45,800
243	711	1182	-18.7	23,800
244	1855	621	-0.6	48,000
245	1189	474	-8.9	59,300
246	551	459	-25.1	61,000
247	1348	604	-6.9	49,100
248	460	448	-29.3	62,100
249	1733	451	-1.9	61,800
250	1974	788	>0.0	39,200
251	808	392	-16.1	69,500
252	874	553	-14.6	52,500
253	753	848	-17.6	36,500
254	995	450	-12.1	61,900
255	1690	679	-2.4	44,600
256	994	1006	-12.1	30,200
257	508	464	-27.4	60,400
258	1517	820	-4.4	37,800

a) Master table of proteins in the rat liver database, showing spot master number, gel position (x and y), isoelectric point relative to CPK standards, and predicted molecular mass (from the standard curve of Fig. 8).

Electrophoresis 1991, 12, 907–930

MSN	X	Y	CPKol	SDSMW
259	1796	961	-1.1	31,900
260	661	1361	-20.4	17,700
261	1725	679	-2.0	44,600
262	496	1127	-28.0	25,800
263	1063	172	-10.9	177,400
265	1390	673	-6.3	45,000
266	510	437	-27.3	63,400
267	660	1038	-20.4	29,000
268	430	961	-31.0	31,900
269	1044	606	-11.2	48,900
270	2019	853	>0.0	36,300
271	857	422	-15.0	65,200
272	895	968	-14.2	31,700
274	1292	712	-7.6	42,900
275	1350	590	-6.9	49,900
276	1670	1089	-2.6	27,100
277	688	538	-19.4	53,700
278	961	718	-13.0	42,600
279	879	570	-14.5	51,300
281	1848	1084	-0.7	27,300
282	1505	525	-4.6	54,800
283	1313	1147	-7.3	25,100
284	1314	829	-7.3	37,400
285	1332	408	-7.1	67,200
286	1277	652	-7.8	46,100
288	1391	824	-6.3	37,600
289	1147	579	-9.5	50,700
290	925	511	-13.6	55,900
291	787	1476	-16.6	13,900
292	1462	818	-5.1	37,800
293	531	449	-26.3	62,000
294	860	698	-14.9	43,600
295	1162	609	-9.3	48,700
296	218	814	<-35.0	38,000
297	1377	979	-6.5	31,300
299	913	1523	-13.9	12,400
300	2012	667	>0.0	45,300
301	702	178	-19.0	169,200
302	494	1280	-28.1	20,400
303	403	1008	-32.6	30,100
304	1843	1585	-0.7	10,300
305	1049	593	-11.1	49,800
306	1608	989	-3.3	30,900
307	1219	916	-8.5	33,700
308	1627	755	-3.0	40,700
309	1524	892	-4.4	34,700
310	1769	1028	-1.5	29,400
311	1609	1451	-3.3	14,700
312	266	1408	<-35.0	16,100
313	1902	1365	-0.3	17,600
314	1316	1395	-7.3	16,600
315	1341	523	-7.0	54,900
318	1104	1053	-10.1	28,500
320	1480	1459	-4.9	14,400
321	850	603	-15.1	49,100
322	1454	1494	-5.3	13,300
323	670	626	-20.0	47,700
324	655	101	-20.6	420,500
325	1521	675	-4.4	44,800
326	1587	677	-3.6	44,700
327	1388	409	-6.3	67,000
328	448	1291	-30.0	20,100
330	1608	751	-3.3	40,900
331	1566	697	-3.8	43,700
332	531	471	-26.3	59,600
333	784	1156	-16.7	24,700
334	1059	407	-10.9	67,300
335	1593	303	-3.5	88,500
336	1616	598	-3.2	49,400
338	1854	1004	-0.6	30,300
339	1265	888	-8.0	34,900
340	581	585	-23.6	50,300
341	1497	1047	-4.7	28,700
343	1351	265	-6.8	102,200
344	1813	549	-0.9	52,800

MSN	X	Y	CPKol	SDSMW
345	1006	578	-11.9	50,800
346	1095	640	-10.3	46,800
347	625	728	-21.7	42,000
348	361	983	-35.3	31,100
349	110	1343	<-35.0	18,300
350	521	1130	-26.7	25,700
351	912	619	-13.9	48,100
352	1574	530	-3.7	54,300
353	961	912	-12.9	33,900
354	706	762	-18.9	40,400
355	1450	830	-5.3	37,300
356	1374	1152	-6.5	24,900
357	474	997	-28.7	30,600
358	798	346	-16.3	77,800
359	764	338	-17.3	79,400
360	1384	1068	-6.4	27,900
361	1713	769	-2.1	40,100
362	1161	859	-9.3	36,100
363	914	1156	-13.8	24,800
364	412	435	-32.0	63,700
365	741	486	-17.9	58,200
366	878	1503	-14.6	13,000
367	1560	935	-3.9	33,000
368	983	520	-12.4	55,200
369	434	441	-31.0	63,000
370	639	610	-21.2	48,700
371	1587	860	-3.6	36,100
372	1875	762	-0.5	40,400
373	1351	1059	-6.8	28,300
374	1506	715	-4.6	42,700
375	1823	532	-0.9	54,200
376	254	417	<-35.0	65,900
377	1409	583	-6.1	50,400
378	621	494	-21.8	57,500
379	1017	595	-11.7	49,600
381	953	598	-13.1	49,400
382	856	674	-15.0	44,900
383	1252	258	-8.1	105,300
384	1699	1518	-2.3	12,500
385	1042	493	-11.2	57,500
386	1490	583	-4.7	50,400
387	1554	603	-4.0	49,100
388	1193	404	-8.9	67,700
389	1374	902	-6.5	34,300
390	1456	969	-5.2	31,700
391	718	690	-18.5	44,000
392	1799	732	-1.1	41,900
393	1482	758	-4.8	40,600
394	1227	1461	-8.4	14,400
395	1530	577	-4.3	50,800
396	1410	755	-6.0	40,800
397	912	256	-13.9	106,400
399	1465	1063	-5.0	28,100
400	1473	450	-4.9	61,900
401	1029	1140	-11.5	25,300
403	1516	754	-4.4	40,800
404	1495	554	-4.7	52,500
405	1525	1092	-4.3	27,100
406	723	252	-18.4	108,000
409	650	663	-20.8	45,500
410	1501	478	-4.6	59,000
411	936	1057	-13.4	28,300
412	350	1120	-35.9	26,000
413	1033	538	-11.4	53,700
415	737	425	-18.0	64,900
416	1578	606	-3.7	48,900
417	646	496	-21.0	57,300
418	1695	482	-2.3	58,600
419	725	770	-18.3	40,000
420	1289	1041	-7.7	28,900
421	1171	912	-9.1	33,900
422	599	162	-22.8	193,700
423	929	856	-13.6	36,200
424	739	625	-17.9	47,700
425	1490	965	-4.7	31,800

MSN	X	Y	CPKol	SDSMW
426	1296	704	-7.6	43,300
427	810	843	-16.0	36,800
428	1565	303	-3.9	88,700
429	1259	847	-8.0	36,600
430	1253	562	-8.1	51,900
431	734	1426	-18.1	15,500
432	483	433	-28.5	63,900
434	518	1041	-26.9	28,900
435	1020	1170	-11.6	24,300
436	1122	196	-9.8	147,600
437	1870	673	-0.5	45,000
438	435	1102	-31.0	26,700
439	86	847	<-35.0	36,600
440	1740	544	-1.8	53,200
441	599	1571	-22.8	10,800
443	743	335	-17.8	80,100
446	801	668	-16.2	45,200
447	1050	926	-11.1	33,300
448	1245	1298	-8.2	19,800
449	1576	1516	-3.7	12,600
450	1818	1021	-0.9	29,600
451	1094	440	-10.3	63,100
452	1945	802	>0.0	38,600
453	1652	894	-2.8	34,600
454	1403	500	-6.1	56,900
456	1394	718	-6.3	42,600
457	905	436	-14.0	63,500
459	1038	581	-11.3	50,500
460	1598	294	-3.4	91,400
461	1528	863	-4.3	35,900
462	1098	1137	-10.2	25,400
463	849	1125	-15.2	25,800
464	1814	1072	-0.9	27,800
465	1388	481	-6.3	58,700
466	1194	1084	-8.9	27,300
468	577	467	-23.9	60,100
469	1140	888	-9.6	34,900
470	1797	524	-1.1	54,800
471	1293	1133	-7.6	25,500
472	618	655	-21.9	46,000
473	2009	299	>0.0	89,900
474	1205	215	-8.7	131,300
475	1035	788	-11.4	39,200
476	160	155	<-35.0	207,600
477	469	1370	-28.9	17,400
478	599	662	-22.8	45,600
479	1009	540	-11.8	53,500
480	1216	235	-8.6	117,400
482	816	346	-15.9	77,800
483	683	673	-19.3	44,900
485	1608	1013	-3.3	30,000
486	478	599	-28.6	49,300
487	1025	607	-11.5	48,800
488	1045	1186	-11.2	23,700
489	1609	301	-3.3	89,200
490	775	1289	-17.0	20,100
491	692	178	-19.3	169,300
492	1100	964	-10.2	31,800
493	1760	776	-1.6	39,700
494	882	247	-14.5	110,700
495	470	1258	-28.9	21,200
496	494	1436	-28.1	15,200
497	980	852	-12.5	36,400
499	1414	546	-6.0	53,100
500	1234	1072	-8.3	27,800
501	1246	659	-8.2	45,700
502	824	792	-15.7	39,000
503	1246	1134	-8.2	25,500
504	1115	1407	-9.9	16,200
505	1189	391	-8.9	69,700
506	1578	402	-3.7	68,000
507	787	250	-16.6	109,000
508	979	552	-12.5	52,800
509	1153	619	-9.4	48,100
510	1730	1006	-2.0	30,200

...esis 199...

X
809
1099
1696
948
481
1334
868
798
822
632
1332
603
1190
479
768
747
1170
1502
1728
507
870
1347
1513
308
1851
1463
909
625
1164
803
1259
856
803
1162
128
1355
595
1369
992
1125
705
1477
980
700
1028
898
789
777
980
1519
1212
760
618
1142
532
771
1068
822
914
1064
1524
1392
982
1487
758
687
930
1888
642
1317
65
1014
732
1627
1009

MSN	X	Y	CPKol	SDSMW
511	809	484	-16.0	58,400
512	1099	533	-10.2	54,100
513	1696	1034	-2.3	29,200
514	948	636	-13.2	47,100
515	481	543	-28.5	53,400
516	1334	1044	-7.1	28,800
517	868	1021	-14.8	29,700
518	798	779	-16.3	39,600
519	822	670	-15.7	45,100
520	632	165	-21.5	189,000
521	1332	830	-7.1	37,300
522	603	1104	-22.6	26,600
523	1190	309	-8.9	86,800
524	479	1226	-28.6	22,300
525	768	1066	-17.2	28,000
526	747	1016	-17.7	29,800
527	1170	231	-9.2	119,600
528	1502	542	-4.6	53,400
530	1728	620	-2.0	48,000
532	507	1011	-27.4	30,000
533	870	489	-14.7	57,900
534	1347	1085	-6.9	27,300
535	1513	346	-4.5	77,800
536	308	654	<-35.0	46,000
538	1851	689	-0.7	44,100
539	1463	982	-5.1	31,100
540	909	561	-13.9	52,000
541	625	289	-21.7	93,100
542	1164	198	-9.2	146,200
543	803	655	-16.2	45,900
544	1259	1143	-8.0	25,200
545	856	1526	-15.0	12,200
546	803	1071	-16.2	27,800
547	1162	274	-9.3	98,400
548	128	1321	<-35.0	19,000
549	1355	1122	-6.8	25,900
550	595	866	-23.0	35,800
552	1369	494	-6.6	57,500
553	992	405	-12.2	67,600
555	1125	410	-9.8	66,900
556	705	975	-18.9	31,400
557	1477	1030	-4.9	29,300
558	980	583	-12.5	50,400
559	700	1109	-19.1	26,400
560	1028	621	-11.5	48,000
562	898	794	-14.1	38,900
564	789	1446	-16.6	14,900
565	777	766	-16.9	40,200
566	980	328	-12.5	81,900
567	1519	611	-4.4	48,600
569	1212	661	-8.6	45,600
570	760	594	-17.4	49,700
571	618	956	-21.9	32,100
573	1142	771	-9.6	40,000
574	532	787	-26.2	39,300
575	771	250	-17.1	109,200
576	1068	534	-10.8	54,100
577	822	734	-15.7	41,800
578	914	754	-13.8	40,800
579	1064	794	-10.8	38,900
580	1524	714	-4.4	42,800
581	1392	783	-6.3	39,400
582	982	686	-12.4	44,200
584	1487	672	-4.8	45,000
585	758	731	-17.4	41,900
586	687	1152	-19.5	24,900
587	930	523	-13.5	55,000
588	1888	774	-0.4	39,900
589	642	485	-21.1	58,300
590	1317	519	-7.3	55,300
591	65	1548	<-35.0	11,500
592	1014	614	-11.7	48,400
593	732	176	-18.1	172,300
594	1627	478	-3.0	59,000
595	1009	1426	-11.8	15,500
596	619	269	-21.9	100,500
597	1176	461	-9.1	60,700
598	1465	1044	-5.0	28,800
599	741	1188	-17.9	23,600
600	907	402	-14.0	68,000
601	687	658	-19.5	45,800
602	712	1138	-18.7	25,400
603	898	181	-14.1	165,200
604	783	1461	-16.7	14,400
605	736	223	-18.0	125,300
606	629	273	-21.6	98,700
607	1064	286	-10.8	94,000
608	883	503	-14.5	56,700
609	2012	610	>0.0	48,700
610	1255	903	-8.1	34,200
612	1103	391	-10.1	69,600
613	778	265	-16.9	102,000
614	824	518	-15.7	55,400
615	1095	195	-10.3	149,100
616	1759	478	-1.6	59,000
617	994	372	-12.1	72,900
618	751	374	-17.6	72,400
619	1429	518	-5.7	55,300
620	1050	520	-11.1	55,200
621	923	1105	-13.7	26,600
622	1462	622	-5.1	47,900
623	759	225	-17.4	124,000
624	758	1038	-17.4	29,000
625	1438	606	-5.5	48,900
626	1096	1089	-10.2	27,200
627	942	548	-13.3	53,000
628	809	621	-16.0	48,000
629	899	979	-14.1	31,300
630	1135	1321	-9.6	19,100
631	979	615	-12.5	48,300
632	1542	1076	-4.1	27,600
633	1345	814	-6.9	38,000
634	409	950	-32.2	32,400
635	1165	704	-9.2	43,300
636	774	604	-17.0	49,000
637	1263	524	-8.0	54,800
638	952	411	-13.1	66,700
639	1717	575	-2.1	51,000
640	994	292	-12.1	92,000
641	165	1224	<-35.0	22,400
642	803	251	-16.2	108,900
643	719	296	-18.5	90,700
644	1100	294	-10.2	91,400
645	534	1263	-26.1	21,000
646	1153	1038	-9.4	29,000
648	1246	204	-8.2	140,000
649	14	1406	<-35.0	16,200
650	1713	1049	-2.1	28,600
651	1986	1183	>0.0	23,800
652	1378	816	-6.5	38,000
653	1442	1165	-5.5	24,400
654	650	806	-20.8	38,400
655	1111	551	-10.0	52,700
656	1095	861	-10.3	36,000
657	1524	540	-4.4	53,600
658	1777	860	-1.4	36,000
659	391	584	-33.4	50,400
660	977	565	-12.5	51,700
661	658	166	-20.5	187,500
662	732	312	-18.1	86,100
663	1787	567	-1.2	51,500
664	888	268	-14.4	100,900
665	889	775	-14.3	39,800
666	715	221	-18.6	126,300
667	781	227	-16.8	122,400
668	646	165	-21.0	189,100
669	1116	353	-9.9	76,300
670	1382	643	-6.4	46,600
671	547	789	-25.3	39,200
673	984	746	-12.4	41,200
674	1661	448	-2.7	62,100
675	1523	562	-4.4	51,900
676	708	642	-18.8	46,700
677	919	615	-13.7	48,300
678	1085	551	-10.5	52,700
679	600	923	-22.7	33,400
680	1237	1004	-8.3	30,300
681	1103	283	-10.1	95,100
682	1406	477	-6.1	59,100
683	1596	249	-3.4	109,800
684	555	699	-24.8	43,500
685	1167	1313	-9.2	19,300
686	1932	790	0.0	39,100
687	1545	619	-4.1	48,100
688	1456	764	-5.2	40,300
689	1011	953	-11.8	32,300
690	1995	270	>0.0	100,200
691	812	888	-16.0	34,900
692	1154	1461	-9.4	14,400
693	1993	819	>0.0	37,800
694	1628	656	-3.0	45,900
695	928	254	-13.6	107,000
696	1854	715	-0.6	42,700
697	1997	345	>0.0	78,000
698	957	563	-13.0	51,800
699	1540	730	-4.2	42,000
702	577	900	-23.8	34,400
703	1610	562	-3.2	51,900
705	1278	571	-7.8	51,200
706	1841	704	-0.7	43,300
707	1018	1386	-11.7	16,900
709	1074	1145	-10.7	25,100
710	293	889	<-35.0	34,800
712	720	412	-18.5	66,600
713	1386	841	-6.4	36,800
714	1328	263	-7.1	103,100
715	698	433	-19.1	63,900
716	701	481	-19.0	58,700
717	1875	699	-0.5	43,600
718	575	702	-23.9	43,400
719	1216	204	-8.6	140,400
721	1069	464	-10.8	60,400
722	1272	506	-7.9	56,400
723	958	822	-13.0	37,700
724	763	395	-17.3	69,100
725	720	916	-18.5	33,700
726	1476	415	-4.9	66,200
727	1846	473	-0.7	59,400
728	510	783	-27.3	39,400
729	1217	1126	-8.6	25,800
730	1858	724	-0.6	42,300
731	665	765	-20.2	40,300
733	1321	312	-7.2	85,900
734	719	427	-18.5	64,600
735	1101	473	-10.2	59,500
736	1359	569	-6.7	51,400
738	696	220	-19.2	127,600
739	687	409	-19.5	67,000
740	1205	256	-8.7	106,200
741	995	563	-12.1	51,900
742	898	596	-14.1	49,500
743	881	181	-14.5	165,900
744	1951	686	>0.0	44,200
745	726	168	-18.3	183,600
746	999	643	-12.0	46,600
748	182	1503	<-35.0	13,000
749	2005	649	>0.0	46,300
750	1448	575	-5.4	51,000
751	792	266	-16.5	101,900
752	469	296	-28.9	90,600
754	664	254	-20.3	107,000
755	1195	184	-8.8	161,000
756	1821	1113	-0.9	26,300
757	909	246	-13.9	111,000
760	790	133	-16.5	264,900

MSN	X	Y	CPKpI	SDSMW
761	1399	733	-6.2	41,800
763	1416	1085	-5.9	27,300
764	2020	569	>0.0	51,400
765	651	475	-20.8	59,300
766	1052	1149	-11.1	25,000
767	1968	468	>0.0	59,900
768	1330	685	-7.1	44,300
769	1970	613	>0.0	48,500
770	857	617	-15.0	48,200
771	1337	974	-7.0	31,500
773	1576	502	-3.7	56,700
775	969	824	-12.8	37,600
776	1438	708	-5.5	43,100
777	1539	458	-4.2	61,000
778	850	434	-15.1	63,800
779	700	411	-19.1	66,800
780	1052	1136	-11.1	25,500
784	1413	529	-6.0	54,400
785	1364	885	-6.7	35,000
786	1822	835	-0.9	37,100
787	893	392	-14.3	69,500
790	616	882	-22.0	35,100
791	451	1429	-29.8	15,400
792	777	377	-16.9	72,000
793	1536	1543	-4.2	11,700
794	1461	807	-5.1	38,300
796	388	546	-33.6	53,100
797	1126	212	-9.8	133,700
798	933	437	-13.5	63,400
799	1420	593	-5.9	49,800
800	1759	279	-1.6	96,500
801	624	865	-21.7	35,800
802	898	547	-14.2	53,000
803	1775	1468	-1.4	14,200
804	573	196	-24.0	148,400
805	203	494	<-35.0	57,400
806	980	1039	-12.5	29,000
807	902	308	-14.1	87,200
808	625	827	-21.7	37,500
809	1851	1015	-0.7	29,900
810	440	573	-30.9	51,100
811	1358	249	-6.8	109,700
812	851	393	-15.1	69,400
813	745	1246	-17.8	21,600
814	2028	810	>0.0	38,200
815	1086	645	-10.4	46,500
816	629	313	-21.6	85,700
817	1376	1177	-6.5	24,000
818	1771	790	-1.4	39,100
819	1045	263	-11.2	103,100
820	984	362	-12.4	74,600
821	1712	279	-2.2	96,700
822	1256	205	-8.1	139,200
823	1517	654	-4.4	46,000
824	1442	449	-5.5	62,000
825	1240	513	-8.3	55,800
826	1309	1014	-7.4	29,900
827	2012	708	>0.0	43,100
828	937	1405	-13.4	16,200
830	1342	756	-7.0	40,700
831	562	826	-24.5	37,500
832	1073	1039	-10.7	29,000
833	481	820	-28.5	37,800
834	501	581	-27.8	50,500
837	751	748	-17.6	41,100
838	635	833	-21.3	37,200
839	1494	459	-4.7	60,900
840	1952	301	>0.0	89,300
841	1585	1080	-3.6	27,500
842	571	1312	-24.1	19,400
843	1325	649	-7.2	46,300
844	1727	301	-2.0	89,200
845	630	679	-21.5	44,600
846	2016	905	>0.0	34,200
847	673	1200	-19.9	23,200

MSN	X	Y	CPKpI	SDSMW
848	1863	271	-0.6	99,500
849	1166	523	-9.2	54,900
850	1535	1024	-4.2	29,600
851	1035	826	-11.4	37,500
852	834	542	-15.5	53,400
855	499	220	-27.8	127,100
856	1063	194	-10.9	150,500
857	887	890	-14.4	34,800
858	1448	639	-5.4	46,900
859	706	311	-18.9	86,200
860	1070	1066	-10.7	28,000
861	472	347	-28.8	77,600
862	674	480	-19.9	58,800
864	1307	499	-7.4	57,000
865	645	887	-21.0	34,900
866	827	1004	-15.6	30,300
868	685	494	-19.5	57,400
869	1807	402	-1.0	68,000
870	1323	783	-7.2	39,400
871	1228	1031	-8.4	29,300
872	1904	346	-0.3	77,700
873	556	647	-24.8	46,400
874	1540	756	-4.2	40,700
875	1566	777	-3.8	39,700
876	1198	351	-8.8	76,800
877	1076	720	-10.6	42,500
878	1161	1111	-9.3	26,400
879	647	757	-20.9	40,700
880	1756	594	-1.6	49,700
881	1543	278	-4.1	97,100
883	1432	890	-5.7	34,800
884	922	689	-13.7	44,100
885	1103	414	-10.1	66,400
886	1501	607	-4.6	48,900
887	798	1103	-16.3	26,600
888	636	634	-21.3	47,200
889	951	759	-13.1	40,600
890	717	548	-18.6	52,900
891	1123	229	-9.8	121,200
892	891	413	-14.3	66,400
894	1245	234	-8.2	117,800
895	1962	346	>0.0	77,700
896	1322	626	-7.2	47,700
897	420	570	-31.4	51,300
898	662	428	-20.3	64,500
899	845	243	-15.3	113,000
900	624	703	-21.7	43,400
901	931	1094	-13.5	27,000
903	799	229	-16.3	121,000
904	765	520	-17.2	55,200
905	775	889	-17.0	34,800
907	888	824	-14.4	37,600
908	828	1303	-15.6	19,700
910	681	1544	-19.7	11,700
911	1544	301	-4.1	89,100
913	1606	387	-3.3	70,400
914	1237	688	-8.3	44,100
916	1442	749	-5.5	41,100
917	1260	367	-8.0	73,700
919	764	1541	-17.3	11,700
920	1133	1123	-9.7	25,900
921	1123	380	-9.8	71,500
923	829	242	-15.6	113,200
924	1131	318	-9.7	84,300
925	1441	874	-5.5	35,400
926	679	219	-19.7	128,200
927	1487	1191	-4.8	23,500
928	1082	775	-10.5	39,800
929	1231	816	-8.4	38,000
931	1609	670	-3.3	45,100
932	810	900	-16.0	34,400
933	965	520	-12.8	55,100
934	947	462	-13.2	60,600
936	865	843	-14.8	36,800
937	1421	1056	-5.9	28,400

MSN	X	Y	CPKpI	SDSMW
939	1197	827	-8.8	37,500
941	1765	885	-1.5	35,000
942	602	472	-22.7	59,600
943	312	498	<-35.0	57,100
944	993	491	-12.1	57,700
945	1300	269	-7.5	100,300
946	630	423	-21.6	65,100
947	187	736	<-35.0	41,600
948	1380	344	-6.5	78,200
949	1766	665	-1.5	45,400
950	1038	193	-11.3	151,000
951	860	152	-14.9	213,000
952	957	701	-13.0	43,400
954	503	547	-27.6	53,000
955	1938	712	>0.0	42,900
957	1010	816	-11.8	37,900
959	768	174	-17.2	174,900
960	596	419	-23.0	65,700
961	557	409	-24.8	67,100
962	887	320	-14.4	83,900
963	564	334	-24.5	80,500
964	969	1155	-12.8	24,800
965	671	255	-20.0	106,600
966	1204	798	-8.7	38,700
967	910	154	-13.9	210,300
968	609	1048	-22.3	28,700
969	1285	206	-7.7	138,900
970	822	232	-15.8	119,300
971	976	437	-12.6	63,400
972	403	567	-32.6	51,600
974	279	495	<-35.0	57,400
975	844	981	-15.3	31,200
976	1124	295	-9.8	91,100
977	994	664	-12.1	45,400
978	1612	642	-3.2	46,700
979	749	1141	-17.7	25,300
980	1064	642	-10.8	46,700
981	1197	911	-8.8	33,900
983	1762	1508	-1.6	12,800
984	1344	317	-6.9	84,700
985	1024	1105	-11.5	26,600
987	739	1159	-17.9	24,600
988	816	555	-15.9	52,400
990	785	361	-16.7	74,900
991	1159	317	-9.3	84,500
992	1090	928	-10.4	33,300
993	1030	701	-11.5	43,400
994	847	811	-15.2	38,200
995	902	461	-14.1	60,700
996	888	847	-14.4	36,600
997	1815	579	-0.9	50,700
998	1205	504	-8.7	56,500
999	617	289	-22.0	93,100
1000	968	290	-12.8	92,700
1001	970	771	-12.7	40,000
1002	1736	478	-1.9	58,900
1003	643	1184	-21.1	23,700
1006	822	487	-15.8	58,100
1007	875	279	-14.6	96,400
1009	291	644	<-35.0	46,600
1010	1386	745	-6.4	41,200
1011	459	541	-29.4	53,500
1012	679	661	-19.7	45,600
1013	1818	1128	-0.9	25,800
1014	1032	634	-11.4	47,200
1015	1629	994	-3.0	30,700
1016	1311	1134	-7.4	25,500
1017	1722	424	-2.0	65,000
1018	1015	743	-11.7	41,300
1020	1574	1219	-3.7	22,500
1021	781	484	-16.8	58,400
1022	1129	83	-9.7	591,300
1023	812	317	-15.9	84,600
1024	785	446	-16.7	62,400
1025	1290	739	-7.7	41,500

MSN	X	Y	CPKpI	SDSMW
1026	405	552	-32.5	52,600
1027	1298	848	-7.5	36,500
1028	856	547	-15.0	53,000
1030	1284	226	-7.7	123,200
1031	986	822	-12.3	37,700
1032	1547	403	-4.1	67,900
1033	1381	551	-6.4	52,700
1034	1525	496	-4.3	57,200
1035	1128	645	-9.7	46,500
1036	1226	274	-8.5	98,300
1039	1761	262	-1.6	103,600
1040	541	839	-25.7	36,900
1041	818	910	-15.8	34,000
1044	1036	485	-11.3	58,300
1045	1439	407	-5.5	67,300
1047	1540	250	-4.2	109,200
1048	1576	635	-3.7	47,100
1049	1089	411	-10.4	66,700
1050	949	1040	-13.2	28,900
1051	426	818	-31.1	37,800
1052	1583	1385	-3.6	16,900
1053	779	1092	-16.8	27,000
1054	1613	620	-3.2	48,000
1055	1380	377	-6.5	72,000
1056	284	663	<-35.0	45,500
1058	1261	746	-8.0	41,200
1060	393	605	-33.3	49,000
1061	1817	645	-0.9	46,600
1062	1245	746	-8.2	41,200
1064	1258	792	-8.1	39,000
1065	705	934	-18.9	33,000
1066	1181	734	-9.0	41,800
1067	529	658	-26.3	45,800
1068	508	696	-27.4	43,700
1069	1898	604	-0.3	49,100
1071	873	609	-14.7	48,700
1073	1768	1128	-1.5	25,800
1075	836	773	-15.4	39,900
1076	1863	861	-0.6	36,000
1078	826	566	-15.7	51,600
1081	971	483	-12.7	58,500
1083	1697	202	-2.3	142,300
1085	1157	794	-9.4	38,900
1090	620	910	-21.9	34,000
1092	1867	597	-0.5	49,500
1093	2019	894	>0.0	34,600
1094	1546	538	-4.1	53,700
1095	1545	477	-4.1	59,100
1098	61	935	<-35.0	33,000
1099	1954	237	>0.0	116,000
1101	588	1048	-23.3	28,600
1102	1050	667	-11.1	45,200
1103	457	797	-29.5	38,800
1105	1884	532	-0.4	54,200
1106	1714	649	-2.1	46,300
1107	1717	546	-2.1	53,100
1108	1976	722	>0.0	42,400
1111	547	1066	-25.3	28,000
1112	1348	621	-6.9	48,000
1115	1385	762	-6.4	40,400
1116	1078	816	-10.6	38,000
1117	975	787	-12.6	39,300
1118	1202	933	-8.7	33,100
1119	1022	1076	-11.6	27,600
1120	1905	616	-0.3	48,300
1121	1512	1301	-4.5	19,700
1122	1114	677	-9.9	44,700
1123	1464	452	-5.1	61,700
1125	1048	857	-11.1	36,200
1126	1122	802	-9.8	38,600
1128	1722	892	-2.1	34,700
1133	1098	825	-10.2	37,500
1139	1830	569	-0.8	51,400
1147	764	1182	-17.3	23,800
1148	1968	724	>0.0	42,300
1153	921	1158	-13.7	24,700
1154	1594	864	-3.5	35,900
1161	637	400	-21.3	68,400
1162	623	397	-21.8	68,800
1163	665	397	-20.2	68,700
1168	564	528	-24.4	54,500
1170	552	529	-25.0	54,500
1171	538	524	-25.9	54,800
1172	545	514	-25.5	55,700
1174	1099	522	-10.2	55,000
1176	1304	586	-7.5	50,200
1177	1366	539	-6.6	53,700
1178	1608	702	-3.3	43,400
1179	1485	224	-4.8	124,900
1180	1459	224	-5.2	124,900
1181	1431	223	-5.7	125,100
1182	1407	223	-6.1	125,200
1183	1383	224	-6.4	124,700
1184	1454	182	-5.3	164,400
1185	1422	183	-5.8	162,600
1186	1394	182	-6.3	164,300
1189	1171	214	-9.2	131,800
1190	1457	286	-5.2	94,200
1191	686	1114	-19.5	26,200
1192	265	893	<-35.0	34,700
1193	403	1292	-32.6	20,000
1194	344	1275	<-35.0	20,600
1195	505	1311	-27.6	19,400
1196	572	1293	-24.1	20,000
1197	639	1502	-21.2	13,000
1198	637	1402	-21.3	16,300
1199	614	1407	-22.1	16,200
1200	637	1431	-21.3	15,400
1201	1095	1394	-10.3	16,600
1202	1719	1545	-2.1	11,600
1203	791	668	-16.5	45,200
1204	964	1021	-12.9	29,700
1205	313	195	<-35.0	148,700
1208	306	194	<-35.0	149,800
1209	320	197	<-35.0	147,400
1210	326	197	<-35.0	146,600
1211	394	294	-33.2	91,400
1212	402	294	-32.7	91,200
1214	386	294	-33.7	91,400
1215	641	329	-21.2	81,600
1216	660	329	-20.4	81,600
1217	914	266	-13.8	101,800
1218	873	245	-14.7	112,000
1219	970	372	-12.7	72,900
1220	1021	298	-11.6	90,100
1221	1392	205	-6.3	139,500
1222	1354	203	-6.8	141,800
1223	1362	205	-6.7	139,500
1224	673	540	-19.9	53,600
1225	614	542	-22.1	53,400
1226	603	539	-22.6	53,600
1227	696	623	-19.2	47,800
1228	707	628	-18.9	47,500
1229	475	447	-28.7	62,300
1230	466	1282	-29.0	20,400
1231	759	1461	-17.4	14,400
1232	1324	1170	-7.2	24,200
1233	1583	1005	-3.6	30,300
1234	1865	809	-0.6	38,200
1235	1812	817	-1.0	37,900
1236	1411	703	-6.0	43,400
1237	1392	682	-6.3	44,500
1238	794	410	-16.4	66,900
1239	769	407	-17.1	67,300
1240	740	406	-17.9	67,500
1241	743	511	-17.8	55,900
1242	713	510	-18.7	56,000
1243	682	509	-19.6	56,100
1244	663	504	-20.3	56,500
1245	565	582	-24.4	50,500
1246	547	577	-25.3	50,800
1247	530	576	-26.3	50,900
1249	516	572	-27.0	51,200
1250	973	536	-12.7	53,900
1251	607	532	-22.4	54,200
1252	665	529	-20.2	54,400
1253	899	766	-14.1	40,200
1254	1311	746	-7.4	41,200
1255	1300	761	-7.5	40,400
1257	1938	712	0.0	42,900
1258	1806	718	-1.0	42,600
1259	1727	715	-2.0	42,700
1260	1629	713	-3.0	42,800
1261	1555	717	-4.0	42,600
1262	1468	717	-5.0	42,600
1263	1413	722	-6.0	42,400
1264	1340	717	-7.0	42,600
1265	1263	717	-8.0	42,600
1266	1182	720	-9.0	42,500
1267	1110	717	-10.0	42,600
1268	1055	717	-11.0	42,600
1269	999	717	-12.0	42,600
1270	959	715	-13.0	42,700
1271	905	712	-14.0	42,900
1272	857	714	-15.0	42,800
1273	810	705	-16.0	43,300
1274	774	711	-17.0	42,900
1277	737	708	-18.0	43,100
1278	702	711	-19.0	42,900
1279	671	710	-20.0	43,000
1280	645	710	-21.0	43,000
1281	617	707	-22.0	43,100
1282	595	704	-23.0	43,300
1283	573	700	-24.0	43,500
1284	552	695	-25.0	43,700
1285	536	694	-26.0	43,800
1286	515	687	-27.0	44,200
1287	496	683	-28.0	44,400
1288	467	669	-29.0	45,200
1289	447	667	-30.9	45,300
1290	427	655	-31.0	45,900
1291	412	655	-32.0	45,900
1292	397	652	-33.0	46,100
1293	381	654	-34.0	46,000
1294	365	653	-35.0	46,100
1295	348	653	<-35.0	46,100

Table 2. Table of some identified proteins

POP name	Protein name	MSN's	Basis for identification
IDS:3_ALPHA_HDDH	3-α-hydroxysteroid-dihydrodiol-dehydrogenase, an enzyme of steroid metabolism	137, 159	Pure protein and antibody provided by Dr. T.M. Penning, Department of Pharmacology, School of Medicine, University of Pennsylvania.
IDS:ACTIN_BETA	β cellular actin, a cytoskeletal protein	38	Homologous position with respect to other mammalian systems
IDS:ACTIN_GAMMA	γ cellular actin, a cytoskeletal protein	68	Homologous position with respect to other mammalian systems
IDS:ALBUMIN	Serum albumin, mature form.	21, 28, 33	Predominance in rat plasma
IDS:APO_A-I	Apo A-I plasma lipoprotein, mature form (tentative).	236, 483	Presence in rat plasma, regulation by some lipid-lowering drugs
IDS:CALMODULIN	Calmodulin, an acidic cytosolic calcium-binding protein	123, 649	Homologous position with respect to other mammalian systems
IDS:CATALASE	Catalase (peroxisomal)	54, 61, 106	Presence in purified peroxisomes, similarity in position to mouse catalase
IDS:CPKSPOTS	Spots contributed by the CPK charge standards (not rat liver proteins)	1257 - 1295	
IDS:CPS	Carbamoyl phosphate synthase	114, 157, 167, 174, 1184, 1185, 1186, 1222	Pure protein provided by Dr. Margaret Marshall, Department of Pharmacology, Medical School, University of Wisconsin - Madison.
IDS:CYTOCHROME_B5	Cytochrome b5	87, 477	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:FABP-L	Liver fatty-acid binding protein	227	Pure protein provided by Dr. Nathan Bass, Department of Medicine, University of California School of Medicine, San Francisco
IDS:HMG-COA_SYNTHASE	Cytosolic HMG-CoA Synthase	133, 144, 235, 413	Antibody provided by Dr. Michael Greenspan, Merck Sharp & Dohme Research Laboratories, Rahway, NJ
IDS:LAMIN_B	Lamin B, a nuclear protein	415, 734	Homologous position with respect to other mammalian systems
IDS:MITCON:1	Mitcon:1 (F1 ATPase β subunit), a mitochondrial inner membrane	17, 49, 71, 340, 1245, 1246, 1247, 1249	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:2	Mitcon:2, a mitochondrial matrix stress protein equivalent to E.	15, 25, 110, 1241, 1242, 1243, 1244	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:MITCON:3	Mitcon:3, a mitochondrial matrix stress protein, likely analog of	18, 35, 226, 600, 1238, 1239, 1240	Homologous position with respect to other mammalian systems, presence in mitochondria
IDS:NADPH_P450_RED	NADPH cytochrome P-450 reductase, frequently co-induced with P-450's	175, 251, 812	Pure protein provided by Dr. Andrew Parkinson, Department of Pharmacology, Toxicology and Therapeutics, University of Kansas Medical Center
IDS:PDI	Protein disulphide isomerase 1	168, 1170, 1171, 1172	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:PLASMA_PROTEINS	Rat plasma proteins observed in liver	21, 28, 33, 44, 72, 102, 115, 197, 236, 246, 248, 257, 293, 332, 347, 364, 369, 419, 432, 463, 468, 518, 562, 605, 623, 666, 667, 725, 738, 790, 865, 903, 926	Plasma coelectrophoresis studies
IDS:PRO-ALBUMIN	Serum albumin precursor	47, 93	Relative position to mature albumin, presence in microsomes
IDS:PYRCARBOX	Pyruvate carboxylase	179, 1180, 1181, 1182, 1183	Pavlica, R.J., et al., BBA (1990) *1022* 115-125.
IDS:SOD	Superoxide dismutase	135	Sequence information obtained by R.M. Van Frank, Lilly Research Laboratories, Indianapolis
IDS:TUBULIN_ALPHA	α tubulin, a cytoskeletal protein	56, 132, 1224, 1252	Homologous position with respect to other mammalian systems
IDS:TUBULIN_BETA	β tubulin, a cytoskeletal protein	50, 1225, 1226, 1251	Homologous position with respect to other mammalian systems

Hb-beta.

Computed : hemoglobi:

Protein

Rabbit r

e 3. Computed pI's of two sets of carbamylated protein standards: Rabbit muscle CPK and human hemoglobin (Hb)

Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	NH2- 7.0	Calc pI	Real CPK
Rabbit muscle CPK	KIRBCM	28	27	17	34	18	1	6.84	0.0
		28	27	17	33	18	1	6.67	-1
		28	27	17	32	18	1	6.54	-2
		28	27	17	31	18	1	6.42	-3
		28	27	17	30	18	1	6.31	-4
		28	27	17	29	18	1	6.21	-5
		28	27	17	28	18	1	6.12	-6
		28	27	17	27	18	1	6.03	-7
		28	27	17	26	18	1	5.94	-8
		28	27	17	25	18	1	5.85	-9
		26	27	17	24	18	1	5.76	-10
		28	27	17	23	18	1	5.67	-11
		28	27	17	22	18	1	5.58	-12
		28	27	17	21	18	1	5.48	-13
		28	27	17	20	18	1	5.39	-14
		28	27	17	19	18	1	5.29	-15
		28	27	17	18	18	1	5.20	-16
		28	27	17	17	18	1	5.12	-17
		28	27	17	16	18	1	5.04	-18
		28	27	17	15	18	1	4.96	-19
		28	27	17	14	18	1	4.89	-20
		28	27	17	13	18	1	4.83	-21
		28	27	17	12	18	1	4.77	-22
		28	27	17	11	18	1	4.71	-23
		28	27	17	10	18	1	4.66	-24
		28	27	17	9	18	1	4.61	-25
		28	27	17	8	18	1	4.56	-26
		28	27	17	7	18	1	4.52	-27
		28	27	17	6	18	1	4.48	-28
		28	27	17	5	18	1	4.44	-29
		28	27	17	4	18	1	4.40	-30
		28	27	17	3	18	1	4.36	-31
		28	27	17	2	18	1	4.32	-32
		28	27	17	1	18	1	4.29	-33
		28	27	17	0	18	1	4.25	-34
		28	27	17	0	18	0	4.22	-35
Hb-beta, human	HBHU	7	8	9	11	3	1	7.18	
		7	8	9	10	3	1	6.79	
		7	8	9	9	3	1	6.53	-1.8
		7	8	9	8	3	1	6.32	-3.2
		7	8	9	7	3	1	6.13	-5.3
		7	8	9	6	3	1	5.96	-7.2
		7	8	9	5	3	1	5.78	-10.0
		7	8	9	4	3	1	5.59	-12.3
		7	8	9	3	3	1	5.37	-15.5
		7	8	9	2	3	1	5.14	-18.0
		7	8	9	1	3	1	4.91	-21.0
		7	8	9	0	3	1	4.71	-25.5
		7	8	9	0	3	0	4.54	-27.2

Electrophoresis 1991, 12, 907–930

Table 4. Computed p*I*'s of some known proteins related to measured CPK p*I*'s

	Protein Name	PIR Name	#ASP 3.9	#GLU 4.1	#HIS 6.0	#LYS 10.8	#ARG 12.5	Calc pI	Real CPK
0	Creatine phospho kinase (CPK), rabbit muscle	KIRBCM	28	27	17	34	18	6.84	0.0
1	Fatty acid-binding protein, rat hepatic	FZRTL	5	13	2	16	2	7.83	-3.0
2	b2-microglobulin, human	MGHUB2	7	8	4	8	5	6.09	-5.0
3	Carbamoyl-phosphate synthase, rat	SYRTCA	72	96	28	95	56	5.97	-5.5
4	Proalbumin (serum albumin precursor), rat	ABRTS	32	57	15	53	27	5.98	-6.2
5	Serum albumin, rat	ABRTS	32	57	15	53	24	5.71	-9.0
6	Superoxid dismutase (Cu-Zn, SOD), rat	A26810	8	11	10	9	4	5.91	-9.2
7	Phospholipase C, phophoinositide-specific (?), rat	A28807	34	42	9	49	21	5.92	-9.2
8	Albumin, human	ABHUS	36	61	16	60	24	5.70	-11.9
9	Apo A-I lipoprotein, rat	A24700	18	24	6	23	12	5.32	-13.7
10	proApo A-I lipoprotein, human	LPHUA1	16	30	6	21	17	5.35	-14.3
11	NADPH cytochrome P-450 reductase, rat	RDRTO4	41	60	21	38	36	5.07	-15.6
12	Retinol binding protein, human	VAHU	18	10	2	10	14	5.04	-16.9
13	Actin beta, rat	ATRTC	23	26	9	19	18	5.06	-17.2
14	Actin gamma, rat	ATRTC	20	29	9	19	18	5.07	-16.8
15	Apo A-I lipoprotein, human	LPHUA1	16	30	5	21	16	5.10	-17.5
16	Apo A-IV lipoprotein, human	LPHUA4	20	49	8	28	24	4.88	-19.7
17	Tubulin alpha, rat	UBRTA	27	37	13	19	21	4.66	-19.8
18	F1ATPase beta, bovine	PWBOB	25	36	9	22	22	4.80	-21.0
19	Tubulin beta, pig	UBPGB	26	36	10	15	22	4.49	-22.5
20	Protein disulphide isomerase (PDI), rat hepatic	ISRTSS	43	51	11	51	9	4.07	-25.0
21	Cytochrome b5, rat	CBRT5	10	15	6	10	4	4.59	-26.0
22	Apo C-II lipoprotein, human	LPHUC2	4	7	0	6	1	4.44	-30.5
	Amino acid pI assumed in calulation:		3.9	4.1	6.0	10.8	12.5		

Electrophoresis 1995, 16, 1977–1981

N. Leigh Anderson[1]
Ricardo Esquer-Blasco[1]
Jean-Paul Hofmann[1]
Lydie Meheus[2]
Jos Raymackers[2]
Sandra Steiner[3]
Frank Witzmann[4]
Norman G. Anderson[1]

[1]Large Scale Biology Corporation, Rockville, MD
[2]Innogenetics NV, Ghent
[3]Sandoz Pharma Ltd, Drug Safety Assessment, Toxicology, Basel
[4]Molecular Anatomy Laboratory, Indiana University Purdue University Columbus, Columbus, IN

An updated two-dimensional gel database of rat liver proteins useful in gene regulation and drug effect studies

We have improved upon the reference two-dimensional (2-D) electrophoretic map of rat liver proteins originally published in 1991 (N. L. Anderson *et al.*, *Electrophoresis* 1991, *12*, 907–930). A total of 53 proteins (102 spots) are now identified, many by microsequencing. In most cases, spots cut from wet, Coomassie Blue stained 2-D gels were submitted to internal tryptic digestion [2], and individual peptides, separated by high-performance liquid chromatography (HPLC), were sequenced using a Perkin-Elmer 477A sequenator. Additional spots were identified using specific antibodies.

Figure 1 shows the current annotated 2-D map of F344 rat liver, analyzed using the Iso-DALT system (20 × 25 cm gels) and BDH 4–8 carrier ampholytes. Both the map itself and the master spot number system remain the same as shown in the original publication. Table 1 lists the important features of each identification shown, including the gel position, p*I*, and M_r for the most abundant or most basic form of each protein. Using this extended base of identified spots, a series of four improved calibration functions has been derived for the p*I* and SDS-M_r axes (the first two of which are shown in Fig. 2A and B). Both forward and reverse functions are derived, so that one can compute the physical properties of a spot with a given gel location, or inversely compute the gel position expected for a protein having given physical properties:

$$Y_{\text{RAT LIVER}} = f_{M_r\text{–RAT LIVER Y}}(M_{r\,\text{SEQUENCE-DERIVED}}) \quad (1)$$

$$X_{\text{RAT LIVER}} = f_{pI\text{–RAT LIVER X}}(pI_{\text{SEQUENCE-DERIVED}}) \quad (2)$$

$$M_{r\,\text{GEL-DERIVED}} = f_{\text{RAT LIVER Y}\to M_r}(Y_{\text{RAT LIVER}}) \quad (3)$$

$$pI_{\text{GEL-DERIVED}} = f_{\text{RAT LIVER X}\to pI}(X_{\text{RAT LIVER}}) \quad (4)$$

A spreadsheet program (in Microsoft Excel) was developed to facilitate flexible computation of p*I*'s from amino acid sequence data, and the results were entered into a relational database (Microsoft Access). A table of spot positions and sequence-derived pI's and M_r's was fitted with a large series of analytic equations using Tablecurve (Jandel Scientific), and the four conversion Eqs. (1)–(4), relating computed p*I* and gel *X* coordinate, or computed molecular weight and gel *Y* coordinate, were selected, based on criteria of simplicity, goodness of fit and favorable asymptotic behavior. Table 2 lists the equations and coefficients. Application of Eqs. (3) and (4) to a spot's *X* and *Y* coordinates, given in [1], produce improved M_r estimates, and allow computation of p*I* directly in pH units, instead of in terms of positions relative to creatine phosphokinase (CPK) charge standards. The inverse Eqs. (1) and (2) were used to compute the gel positions of a series of p*I* and M_r tick marks. These tick marks were plotted with SigmaPlot (Jandel), together with fiducial marks locating several prominent spots, and the resulting graphic was aligned over the synthetic gel image (computed by Kepler from the master gel pattern) using Freelance (Lotus Development). Maps were printed as Postscript output from Freelance, either in black and white (as shown here) or in color, where label color indicates subcellular location (available from the first author upon request). We have also used the rat liver 2-D pattern as presented here to calibrate the patterns of other samples. Using mixtures of rat liver and mouse liver samples, for example, we made composite 2-D patterns that allow use of the rat pattern to standardize both axes of the mouse pattern. This was accomplished by deriving transformations relating the rat and mouse *X*, and separately the rat and mouse *Y*, axes (Table 2, lower half; Fig. 2C and D) based on a series of spots that coelectrophorese in these closely related species. These functions were then applied to derive equations relating the mouse liver *X* and *Y* to p*I* and SDS-M_r (Eqs. 5 and 6 below). The resulting standardized 2-D pattern for B6C3F1 mouse liver is shown in Fig. 3.

$$M_{r\,\text{MOUSE LIVER}} = f_{\text{RAT LIVER Y}\to M_r}(f_{\text{MOUSE LIVER Y}\to\text{RAT LIVER Y}}(Y_{\text{MOUSE LIVER}})) \quad (5)$$

$$pI_{\text{MOUSE LIVER}} = f_{\text{RAT LIVER X}\to pI}(f_{\text{MOUSE LIVER X}\to\text{RAT LIVER X}}(X_{\text{MOUSE LIVER}})) \quad (6)$$

A slightly more complex approach can be used to standardize samples that have few or no spots co-electrophoresing with rat liver proteins. In this case, a 2-D gel is prepared with a mixture of the two samples, and four functions (forward and backward, each for *X* and *Y*) are derived relating each sample's own master pattern to the composite. The required functions are then applied in a nested fashion to yield the desired result (using rat plasma as an example):

$$M_{r\,\text{RAT PLASMA}} = f_{\text{RAT LIVER Y}\to M_r}(f_{\text{RAT PLASMA+LIVER Y}\to\text{RAT LIVER Y}}(f_{\text{RAT PLASMA Y}\to\text{RAT PLASMA+LIVER Y}}(Y_{\text{RAT PLASMA}}))) \quad (7)$$

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Keywords: Two-dimensional polyacrylamide gel electrophoresis / Liver / Map / Identification / Calibration

0173-0835/95/1010-1977 $5.00+.25/0

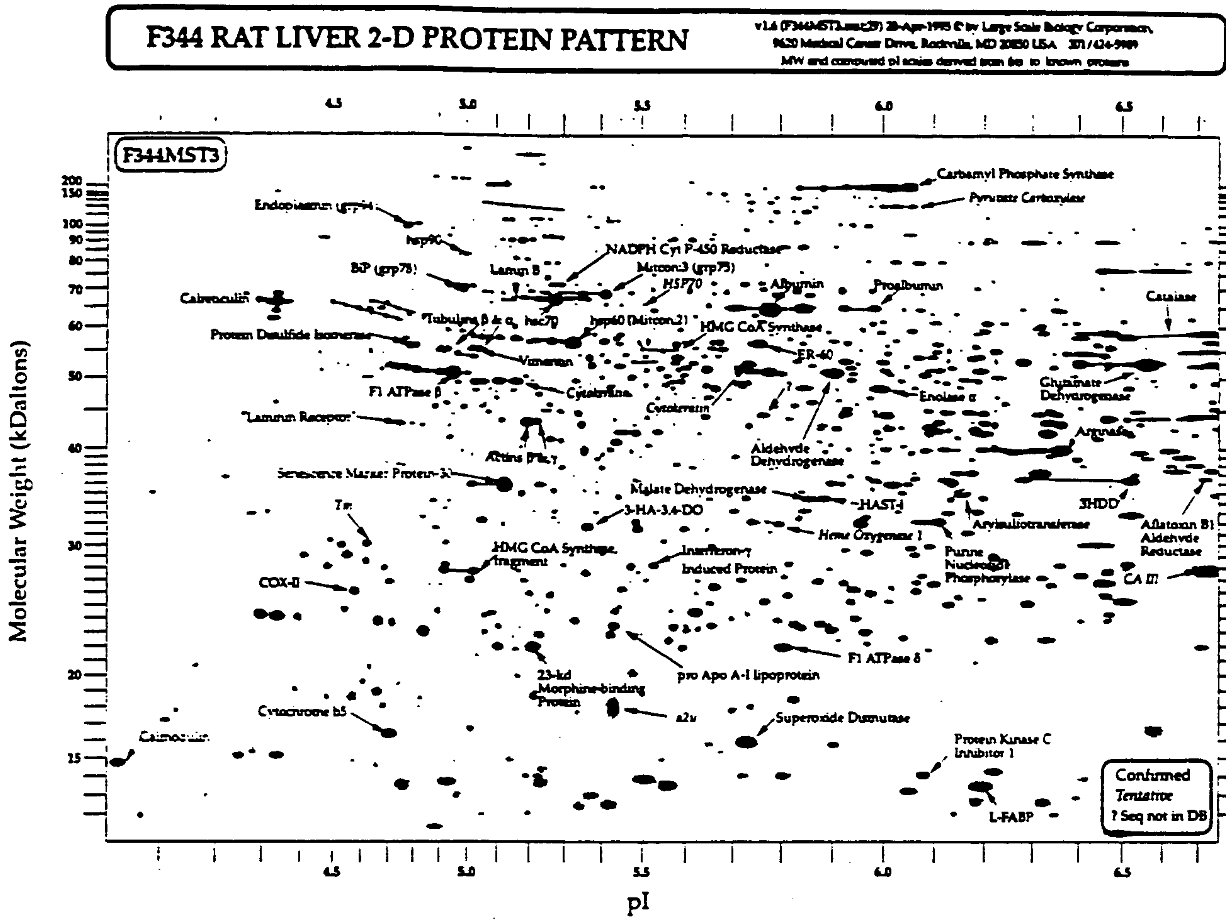


Figure 1. Master 2-D gel pattern of Fischer 344 rat liver proteins, annotated with 53 protein identifications and computed p*I* and M_r axes. Tentative identifications are in italic type.

Table 1. Proteins identified in the 2-D pattern of F344 rat liver

MSN[a)]	Protein ID[b)]	Protein name	Identification comments	Gel X[c)]	Experimental p*I*[d)]	Gel Y[c)]	Experimental M_r[d)]
126	HADO-HUMAN[e)]	3-HA-3,4-DO: 3-hydroxy-anthranilate-3,4-dioxy-genase	Internal sequence	871.95	5.36	921.35	30 207
137, 159, 288, 258	DIDH_RAT	3HDD: 3-hydroxysteroid dihydrodiol reductase	Ab (T.M. Penning) and pure protein	1857.52	6.51	822.52	34 406
173	MUP_RAT	α_2u globulin	Presence in liver microsome lumen, abundance in kidney, p*I*, M_r	919.16	5.43	1313.81	19 549
38	ACTB_HUMAN	Actin β	Analogy with other mammalian patterns (*e.g.* human) through coelectrophoresis	763.40	5.19	693.64	41 586
68	ACTG_HUMAN	Actin γ	Analogy with other mammalian patterns (*e.g.* human) through coelectrophoresis	779.42	5.21	692.26	41 677
693	AFAR_RAT	Aflatoxin B1 aldehyde reductase	Internal sequence	1993.32	6.72	818.60	34 593
28, 21, 33	ALBU_RAT	Albumin	Coelectrophoresis with principal plasma protein	1262.81	5.86	445.64	66 354
43	DHAM_RAT	Aldehyde dehydrogenase	*N*-Terminal sequence and AAA	1317.72	5.91	589.03	49 602
96	ARGI_RAT	Arginase	Internal sequence	1730.72	6.34	756.02	37 819
117	SUAR_RAT	Arylsulfotransferase	Internal sequence	1547.96	6.14	849.08	33 186
1163, 1161, 1162, 20	GR78_RAT	BIP (GRP-78)	Ab (F. Witzmann)	665.33	5.01	397.39	74 564
185	CAH3_RAT	CA-III	Uncertain; by comparison with mouse	1996.60	6.72	1017.02	26 887
123	CALM_HUMAN	Calmodulin	Analogy with human cellular patterns through coelectrophoresis	23.05	4.03	1433.25	17 419
3, 201, 48, 39, 22, 24	CRTC_RAT	Calreticulin	Ab (Lance Pohl)	310.59	4.34	433.80	68 206

Table 1. continued

MSN[a)]	Protein ID[b)]	Protein name	Identification comments	Gel X[c)]	Experimental p*I*[d)]	Gel Y[c)]	Experimental M_r[d)]
1184, 1186, 114, 174, 1185, 167, 157	CPSM_RAT	Carbamyl phosphate synthase	2-D of pure protein; comfirmed by *N*-terminal sequence and AAA	1453.56	6.05	181.64	160 640
54, 61	CATA_RAT	Catalase	Internal sequence	2000.81	6.73	499.64	58 968
136	COX2_RAT	COX-II	Ab (J. W. Taanman), confirmed by internal sequence	452.57	4.61	1062.67	25 504
87	CYB5_RAT	Cytochrome B5	2-D of pure protein; Ab; confirmed by AAA	515.68	4.73	1370.55	18 493
41	CK-RAT[e)]	Cytokeratin	Location in cytoskeletal fraction	1165.12	5.75	569.09	51 448
29	CK-RAT[e)]	Cytokeratin	Location in cytoskeletal fraction	743.11	5.15	605.23	48 187
5, 11	ENPL-RAT[e)]	Endoplasmin	Ab (F. Witzmann)	567.73	4.83	263.37	112 194
60	ENOA_RAT	Enolase A	Internal sequence and AAA	1399.78	6.00	623.54	46 674
27	ER60_RAT	ER-60	*N*-Terminal sequence (R. M. Van Frank)	1184.20	5.77	523.51	56.169
17	ATPB_RAT	F1 ATPase β	*N*-Terminal sequence and AAA	629.06	4.95	588.83	49 620
196	ATP7_RAT	F1 ATPase δ	Internal sequence	1227.24	5.82	1184.65	22 310
79	F16P_RAT	Fructose-1,6-bis-phosphatase	Uncertain; by comparison with ID in Garrison and Wager (JBC 257:13135-13143)	924.54	5.44	737.77	38 858
62, 78	DHE3_RAT	Glutamate dehydrogenase	*N*-Terminal sequence and internal sequence	1887.39	6.55	566.92	51 655
125	HAST-RAT[e)]	HAST-1: N-hydroxyaryl-amine sulfotransferase	Internal sequence	1297.94	5.89	861.55	32 638
307	HO1_RAT	Heme oxygenase 1	Uncertain; available data from internal sequence	1219.39	5.81	915.71	30 423
413, 1250, 933	HMCS_RAT	HMG CoA synthase, cytosolic	Ab (J. Germershausen)	1033.48	5.59	538.13	54 571
133, 144, 235	HMCS_RAT	HMG CoA synthase, mitochondrial (frag)	Ab (J. Germershausen), *N*-terminal sequence (Steiner/Lottspeich)	666.40	5.02	1019.42	26 811
8, 23, 1307	HS7C_RAT	HSC-70	Positional homology (with human, *etc.*) through coelectrophoresis	811.87	5.27	425.76	69 521
15, 25, 110	P60_RAT	HSP-60	Ab (F. Witzman); confirmed by *N*-terminal sequence and AAA	845.09	5.32	520.03	56 561
971	HS70-RAT[e)]	HSP-70	Ab (F. Witzman)	976.11	5.51	437.14	67 674
1216, 1215, 90	HS90-RAT[e)]	HSP-90	Ab (F. Witzman)	659.86	5.00	329	90 107
256	INGI-HUMAN	Interferon-γ induced protein	Internal sequence	993.85	5.54	1006.04	27 237
415, 734	LAMB-RAT[e)]	Lamin B	Positional homology with human through coelectrophoresis, nuclear location	737.10	5.14	425.19	69 615
80	LAMR-RAT[e)]	"Laminin receptor"	Internal sequence	534.02	4.77	697.62	41 327
227	FABL_RAT	L-FABP (liver fatty acid binding protein)	Ab (N. M. Bass)	1586.09	6.18	1483.43	16 622
134	MDHC_MOUSE	Malate dehydrogenase	Internal sequence	1270.85	5.86	861.96	32 620
18, 35, 226	GR75-RAT[e)]	Mitcon:3; grp75	Positional homology with human through coelectrophoresis	905.67	5.41	413.67	71 589
175, 251	NCPR_RAT	NADPH P450 reductase	2-D of pure protein	824.69	5.29	393.21	75 366
1168, 1170, 1171	PDI_RAT	PDI: Protein disulfide isomerase	*N*-Terminal sequence (R. M. van Frank), Ab	564.30	4.83	528.47	55 618
47, 93	ALBU_RAT	Pro-Albumin	Microsomal lumen location, p*I*, M_r relative to albumin	1391.03	5.99	446.68	66 195
236	APA1_RAT	Pro-APO A-I lipoprotein	Coelectrophoresis with plasma protein	920.41	5.43	1137.51	23 467
320	IPK1_BOVIN	Protein kinase C inhibitor 1	Internal sequence; homology with bovine protein	1480.01	6.08	1458.81	17 007
152	PNPH_MOUSE	Purine nucleoside phosphorylase	Internal sequence	1507.19	6.10	911.16	30 599
1179, 1180, 1181, 1182, 1183	PYVC-RAT[e)]	Pyruvate carboxylase	Tentative; 2-D of pure protein (J. G. Henslee, *JBC*, 1979); reported in *Biochim. Biophys. Acta 1022*, 115–125	1485.10	6.08	223.52	131 589
55, 103	SM30_RAT	SMP-30: Senescence marker protein-30	Internal sequence	721.71	5.11	830.10	34 051
135	SODC_RAT	Superoxide dismutase	AAA; comfirmed by internal sequence (R. M. Van Frank)	1161.24	5.74	1388.68	18 173
172	TPM-RAT[e)]	Tm: tropomyosin	Location in cytoskeleton, 2-D position relative to human, Ab	476.24	4.66	957.86	28 865
277, 56	TBA1_RAT	Tubulin α	Positional homology with human through coelectrophoresis, cytoskeletal location	688.22	5.06	537.67	54 620
50, 1225	TBB1_RAT	Tubulin β	Positional homology with human through coelectrophoresis, cytoskeletal location	621.29	4.93	535.48	54 855
1224	VIME_RAT	Vimentin	Positonal homology with human through coelectrophoresis, cytoskeletal location	673.00	5.03	539.50	54 426

Table 1. continued

MSN[a)]	Protein ID[b)]	Protein name	Identification comments	Gel X[c)]	Experimental pI[d)]	Gel Y[c)]	Experimental M_r[d)]
113	Unknown	?: not in sequence databases	Internal sequence	1191.28	5.78	680.42	42 469
104	BBPL_RAT	23 kDa morphine-binding protein	Internal sequence	773.31	5.20	1182.41	22 363

a) Master spot number (MSN) from [1]
b) SwissPROT identifier
c) Coordinates of the most basic or most abundant assigned spot on the F344 master gel pattern
d) pI and M_r of the most basic or most abundant assigned spot, derived from the calibration functions included here
e) SwissPROT style proposed identifier
Abbreviations: AAA, amino acid analysis; Ab, antibody

Table 2. Equations and coefficients

Function	Equation (f)	r2	a	b	c	d	e
Rat gel Y = f(computed M_r)	$y = a + b\exp(-x/c)$	0.988181021	178.74803	1967.7892	32363.958		
Rat gel X = f(computed pI)	$y = a + bx + cx/\ln x + d/x + e/x^{1.5}$	0.99247216	−8685665.5	−904497.94	3856926.1	18276844	−27154534
Computed M_r = f(rat gel Y)	$y = a + bx^c$	0.9960177	−8464.5809	19095881	−0.9086255		
Computed pI = f(rat gel X)	$y = a + bx + cx^2 + dx^2 \ln x + ex^3$	0.99176499	4.044686	−0.00114238	0.0000323	−0.00000455	0.00000000176
Mouse gel Y = f(rat gel Y)	$y = a + bx + cx^{1.5} + dx^{0.5} \ln x + cx/\ln x$	0.99951069	11861.44	678.91666	−0.78964914	1567.5639	−6953.9592
Mouse gel X = f(rat gel X)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99926349	58.935923	0.00091353	−0.000213688	0.00000159	
Rat gel Y = f(mouse gel Y)	$y = a + bx^2 \ln x + cx^{2.5} + dx^3$	0.99950032	69.740526	0.00050772	−0.000130392	0.00000116	
Rat gel X = f(mouse gel X)	$y = a + bx + cx^2 \ln x + dx^{2.5} + ex^3$	0.9992832	−198.07189	2.0899063	−0.000671191	0.000145189	−0.000000986

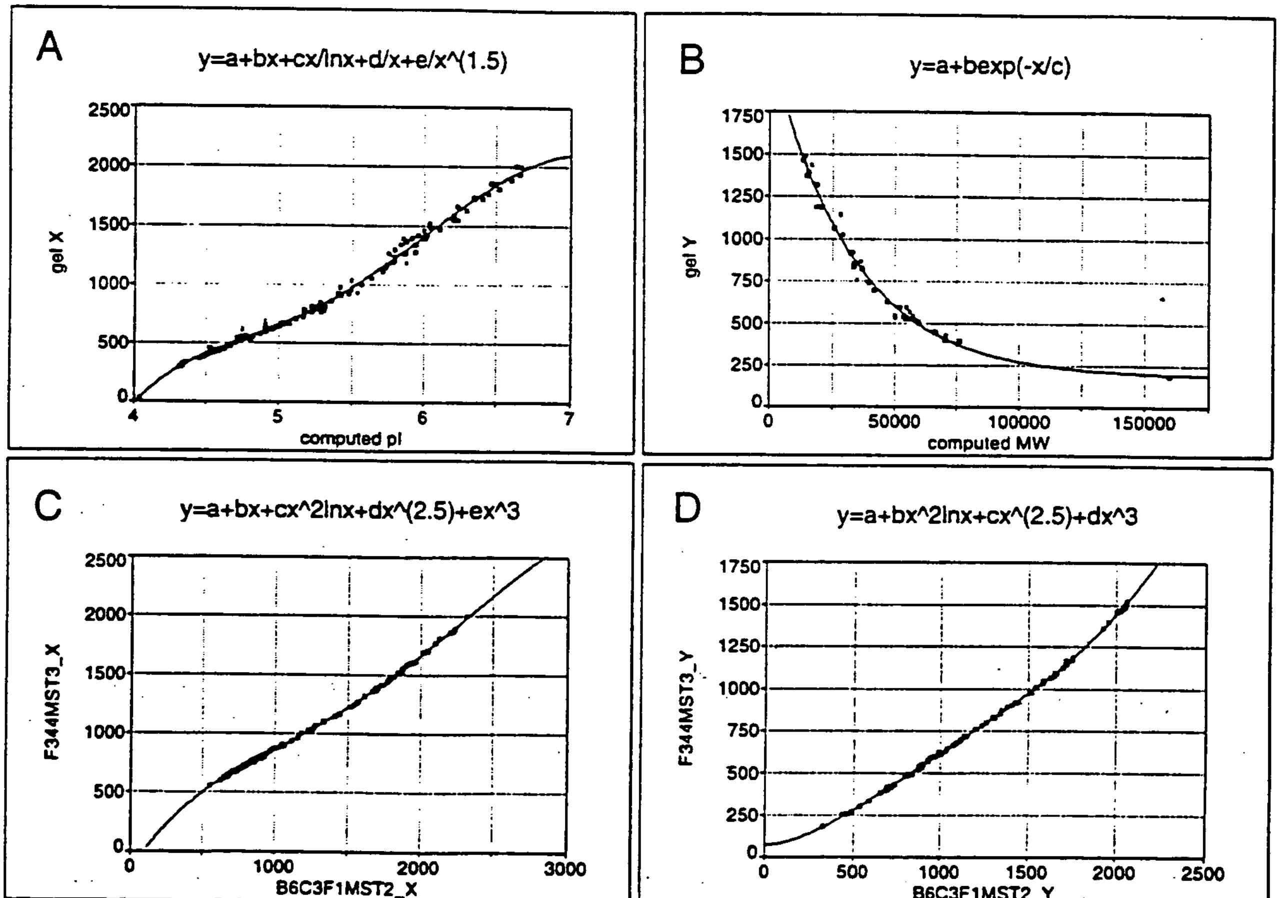


Figure 2. Plots showing fits of selected equations (continuous curves) to data on identified proteins (square symbols). (A) pI computed from sequence data *versus* gel X position for identified spots in F344 rat liver; (B) M_r computed from sequence data *versus* gel Y position for identified spots in F344 rat liver; (C) gel X position for spots in B6C3F1 mouse liver *versus* X position in F3443 rat liver, for coelectrophoresing spots; (D) gel Y position for spots in B6C3F1 mouse liver *versus* Y position in F3443 rat liver, for coelectrophoresing spots. In each case, inverse equations were also computed (Table 2).

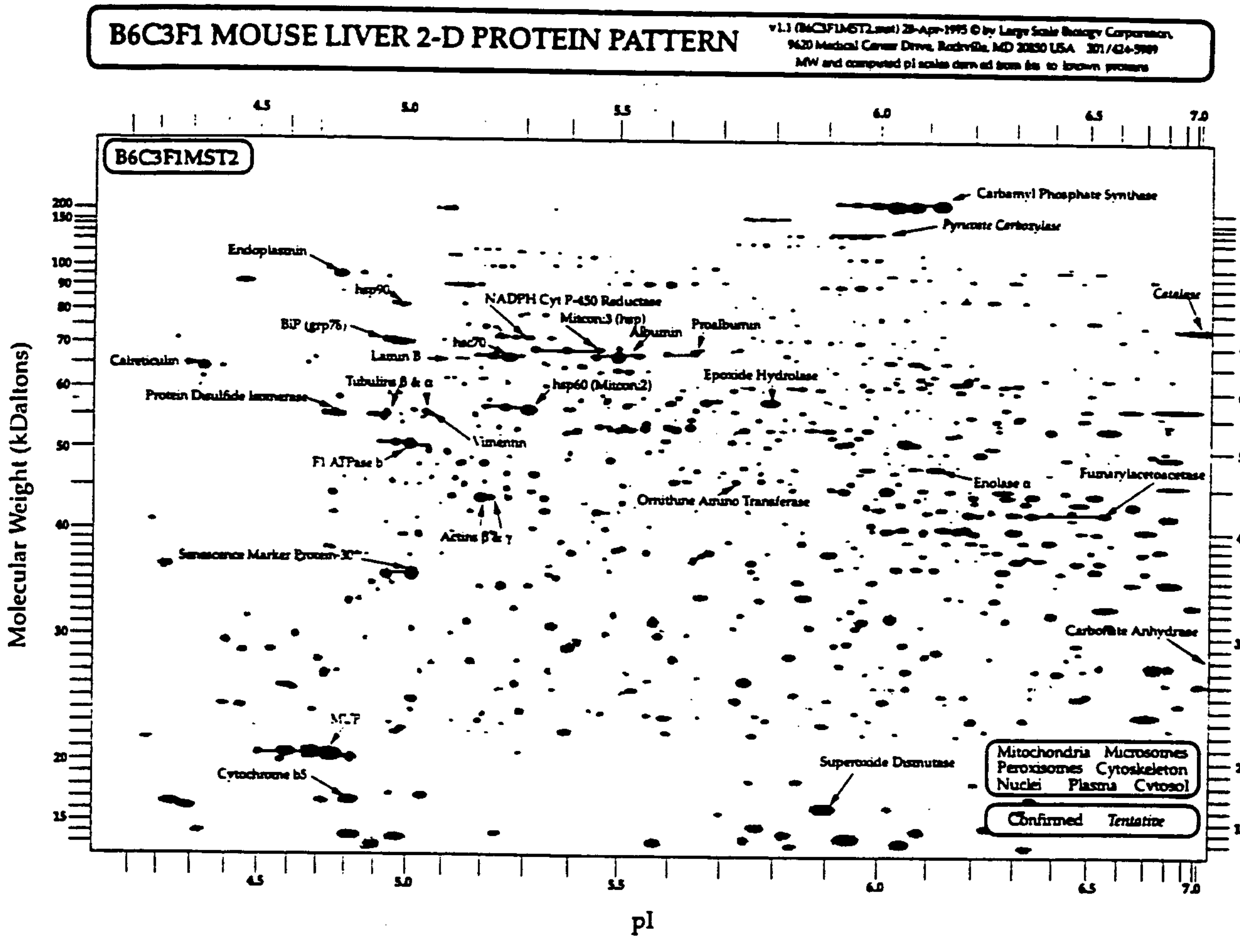


Figure 3. Master 2-D gel pattern for B6C3F1 mouse liver, standardized using the F344 rat liver pattern identifications, according to the method described in the text. Twenty-nine proteins are identified.

$$pI_{RATPLASMA} = f_{RATLIVER\ x \to pI}(f_{RATPLASMA+LIVER\ x \to RATLIVER\ x}(f_{RATPLASMA\ x \to RATPLASMA+LIVER\ x}(X_{RAT\ PLASMA}))) \quad (8)$$

This unified approach, in which one well-populated 2-D pattern is used to standardize a family of other patterns, has the additional advantage that the resulting p*I* and M_r scales are directly compatible. Hence one can compare the relative p*I*'s of mouse and rat versions of a sequenced protein in a consistent p*I* measurement system, and select likely inter-species analogs based on positional relationships on common scales. Adoption of immobilized pH gradient (IPG) technology [4–7] will result in substantial improvements in p*I* positional reproducibility for standard 2-D maps such as those presented here; however, we believe that our approach will continue to be useful in establishing the empirical pH gradient actually achieved by such gels under given experimental conditions (temperature, urea concentration, *etc.*), in relating patterns run on different IPG ranges and using different lots of IPG gels (between which some variation will persist). Development of rodent organ maps is a continuing effort in our laboratories [8–10], and results in regular additions of identified proteins. Those who wish to receive current rodent liver maps, with color annotations, should send a stamped self-addressed envelope to the first author.

We would like to thank the individuals who provided antibodies mentioned in Table 1, and R. M. van Frank for unpublished sequenced data.

Received May 31, 1995

References

[1] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, *12*, 907–930.

[2] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, *203*, 173–179.

[3] Witzmann, F., Clack, J., Fultz, C., Jarnot, B., *Electrophoresis* 1995, *16*, 451–459.

[4] Rosengren, A. E., Bjellqvist, B., Gasparic, V., *US Patent* 4130470, December 1978.

[5] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.

[6] Görg, A., Postel, W., Günther, S., Weser, J., *Electrophoresis* 1985, *6*, 599–604.

[7] Gianazza, E., Astrua-Testori, S., Giacon, P., Righetti, P. G., *Electrophoresis* 1985, *6*, 332–339.

[8] Myers, T. G., Dietz, E. C., Anderson, N. L., Khairallah, E. A., Cohen, S. D., Nelson, S. D., *Chem. Res. Toxicol.* 1995, *8*, 403–413.

[9] Cunningham, M. L., Pippin, L. L., Anderson, N. L., Wenk, M. L., *Toxicol. Appl. Pharmacol.* 1995, *131*, 216–223.

[10] Anderson, N. L., Copple, D. C., Bendele, R. A., Probst, G. S., Richardson, F. C., *Fundam. Appl. Toxicol.* 1992, *18*, 570–580.

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref No. C

Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It

MARC R. WILKINS[1], JEAN-CHARLES SANCHEZ[3], ANDREW. A. GOOLEY[1], RON D. APPEL[3], IAN HUMPHERY-SMITH[2], DENIS F. HOCHSTRASSER[3] AND KEITH L. WILLIAMS[1,*]

[1] *Macquarie University Centre for Analytical Biotechnology, Macquarie University, Sydney, NSW 2109, Australia;* [2] *Department of Microbiology, University of Sydney, NSW, 2006, Australia and* [3] *Central Clinical Chemistry Laboratory and Medical Computing Centre of the University of Geneva, CH 1211 Geneva 14, Switzerland*

Introduction

The advent of large genome sequencing projects has changed the scale of biology. Over a relatively short period of time, we have witnessed the elucidation of the complete nucleotide sequence for bacteriophage λ (Sanger *et al.*, 1982), the nucleotide sequence of an eukaryotic chromosome (Oliver *et al.*, 1992), and in the near future will see the definition of all open reading frames of some simple organisms, including *Mycoplasma pneumoniae, Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans* and *Arabidopsis thaliana*. Nevertheless, genome sequencing projects are not an end in themsleves. In fact, they only represent a starting point to understanding the function of an organism. A great challenge that biologists now face is how the co-expression of thousands of genes can best be examined under physiological and pathophysiological conditions, and how these patterns of expression define an organism.

There are two approaches that can be used to examine gene expression on a large scale. One uses nucleic acid-based technology, the other protein-based technology. The most promising nucleic-acid based technology is differential display of mRNA (Liang and Pardee, 1992; Bauer *et al.*, 1993), which uses polymerase chain reaction with arbitrary primers to generate thousands of cDNA species, each which correspond to an expressed gene or part of a gene. However, it is currently unclear if this technique can be developed to reliably assay the expression of thousands of genes or

* Corresponding Author

identify all cDNA species, and the approach does not easily allow a systematic screening. Analysis of gene expression by the study of proteins present in a cell or tissue presents a favorable alternative. This can be achieved by use of two-dimensional (2-D) gel electrophoresis, quantitative computer image analysis, and protein identification techniques to create 'reference maps' of all detectable proteins. Such reference maps establish patterns of normal and abnormal gene expression in the organism, and allow the examination of some post-translational protein modifications which are functionally important for many proteins. It is possible to screen proteins systematically from reference maps to establish their identities.

To define protein-based gene expression analysis, the concept of the 'proteome' was recently proposed (Wilkins *et al.*, 1995; Wasinger *et al.*, 1995). A proteome is the entire PROTein complement expressed by a genOME, or by a cell or tissue type. The concept of the proteome has some differences from that of the genome, as while there is only one definitive genome of an organism, the proteome is an entity which can change under different conditions, and can be dissimilar in different tissues of a single organism. A proteome nevertheless remains a direct product of a genome. Interestingly, the number of proteins in a proteome can exceed the number of genes present, as protein products expressed by alternative gene splicing or with different post-translational modifications are observed as separate molecules on a 2-D gel. As an extrapolation of the concept of the 'genome project', a 'proteome project' is research which seeks to identify and characterise the proteins present in a cell or tissue and define their patterns of expression.

Proteome projects present challenges of a similar magnitude to that of genome projects. Technically, the 2-D gel electrophoresis must be reproducible and of high resolution, allowing the separation and detection of the thousands of proteins in a cell. Low copy number proteins should be detectable. There should be computer gel image analysis systems that can qualitatively and quantitatively catalog the electrophoretically separated proteins, to form reference maps. A range of rapid and reliable techniques must be available for the identification and characterisation of proteins. As a consequence of a proteome project, protein databases must be assembled that contain reference information about proteins; such databases must be linked to genomic databases and protein reference maps. Databases should be widely accessible and easy to use.

Recently, there have been many changes in the techniques and resources available for the analysis of proteomes. It is the aim of this chapter to discuss the status of the areas outlined above, and to review briefly the progress of some current proteome projects.

Two-dimensional electrophoresis of proteomes

Two dimensional (2-D) gel electrophoresis involves the separation of proteins by their isoelectric point in the first dimension, then separation according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since first described (Klose, 1975; O'Farrell, 1975; Scheele, 1975), it has become the method of choice for the separation of complex mixtures of proteins, albeit with many modifications to the original techniques. 2-D electrophoresis forms the basis of proteome projects through separating proteins by their size and charge (Hochstrasser *et al.*,

HEPG2 2D-PAGE MAP

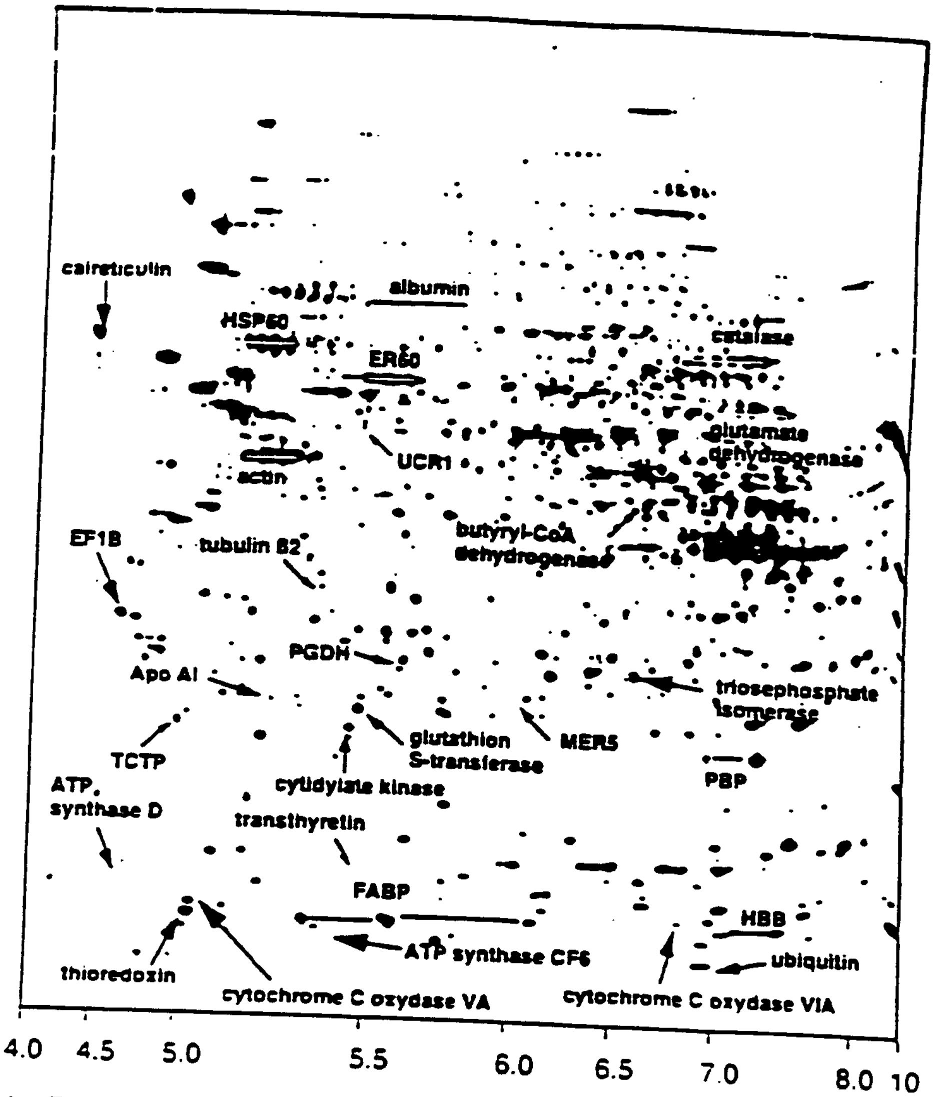


Figure 1. Two-dimensional gel electrophoresis map of a human hepatoblastoma-derived cell line, illustrating the very high resolution of the technique. The first dimensional separation (right to left of figure) was achieved using immobilised pH gradient electrophoresis of 4.0 to 10.0 units. The second dimension (top to bottom of figure) was SDS-PAGE using a 11%–14% acrylamide gradient, allowing separation in the molecular weight range 10–250 kDa. Proteins were visualised by silver staining. Arrows show proteins of known identity.

1992; Celis *et al.*, 1993; Garrels and Franza, 1989; VanBogelen *et al.*, 1992). Current protocols can resolve two to three thousand proteins from a complex sample on a single gel (*Figure 1*).

2-D GEL RESOLUTION AND REPRODUCIBILITY

A primary challenge of separating complex mixtures of proteins by 2-D gel electrophoresis has been to achieve high resolution and reproducibility. High resolution ensures that a maximum of protein species are separated, and high reproducibility is

vital to allow comparison of gels from day to day and between research sites. These factors can be difficult to achieve.

Carrier ampholytes are a common means of isoelectric focusing for the first dimension of 2-D electrophoresis. Gels are usually focused to equilibrium to separate proteins in the pI range 4 to 8, and run in a non-equilibrium mode (NEPHGE) to separate proteins of higher pI (7 to 11.5) (O'Farrell, 1975; O'Farrell, Goodman and O'Farrell, 1977). Unfortunately, the use of carrier ampholytes in the isoelectric focusing procedure is susceptible to 'cathode drift', whereby pH gradients established by prefocusing of ampholytes slowly change with time (Righetti and Drysdale, 1973). Carrier ampholyte pH gradients are also distorted by high salt concentration of samples (Bjellqvist *et al.*, 1982), and by high protein load (O'Farrell, 1975). A further limitation is that iso electric focusing gels, which are cast and subject to electrophoresis in narrow glass tubes, need to be extruded by mechanical means before application to the second dimension – a procedure that potentially distorts the gel. Nevertheless, many of the above shortcomings can be avoided by loading small amounts of ^{14}C or ^{35}S radiolabelled samples (Garrels, 1989; Neidhardt *et al.*, 1989; Vandekerkhove *et al.*, 1990). High sensitivity detection is then achieved through use of fluorography or phosphorimaging plates (Bonner and Laskey, 1974; Johnston, Pickett and Barker, 1990; Patterson and Latter, 1993). However, this approach is only practicable for organisms or tissues that can be radiolabelled.

An alternative technique, which is becoming the method of choice for the first dimension separation of proteins, involves isoelectric focusing in immobilized pH gradient (IPG) gels (Bjellqvist *et al.*, 1982; Görg, Postel and Gunther, 1988; Righetti, 1990). Immobilized pH gradients are formed by the covalent coupling of the pH gradient into an acrylamide matrix, creating a gradient that is completely stable with time. IPG gels are usually poured onto a stiff backing film, which is mechanically strong and provides easy gel handling (Ostergren, Eriksson and Bjellqvist, 1988). The major advantages of IPG separations are that they do not suffer from cathodic drift, they allow focusing of basic and very acidic proteins to equilibrium, pH gradients can be precisely tailored (linear, stepwise, sigmoidal), and that separations over a very narrow pH range are possible (0.05 pH units per cm) (Righetti, 1990; Bjellqvist *et al.*, 1982, 1993a; Sinha *et al.*, 1990; Gorg *et al.*, 1988; Gelfi *et al.*, 1987; Gunther *et al.*, 1988). However, it is not currently possible to use IPG gels to separate very basic proteins of isoelectric point greater than 10, although this is under development. Narrow pH range separations are useful to address problems of protein co-migration in complex samples, allowing 'zooming in' on regions of a gel (*Figure 2*). IPG gel strips are now commercially available, which begin to address the problems of intra- and inter-lab isoelectric focusing reproducibility.

There are two means of electrophoresis for the second dimension separation of proteins: vertical slab gels and horizontal ultrathin gels (Görg, Postel, and Gunther, 1988). Both are usually SDS-containing gradient gels of approximately 11% to 15% acrylamide, which separate proteins in the molecular mass range of 10 – 150kD. A stacking gel is not usually used with slab gels, but is necessary when using horizontal gel setups (Görg, Postel and Gunther, 1988). Comparisons have shown that there is little or no difference in the reproducibility of electrophoresis using either approach (Corbett *et al.*, 1994a), but commercially available vertical or horizontal precast gels will provide greater reproducibility for occasional users. For slab gel electrophoresis,

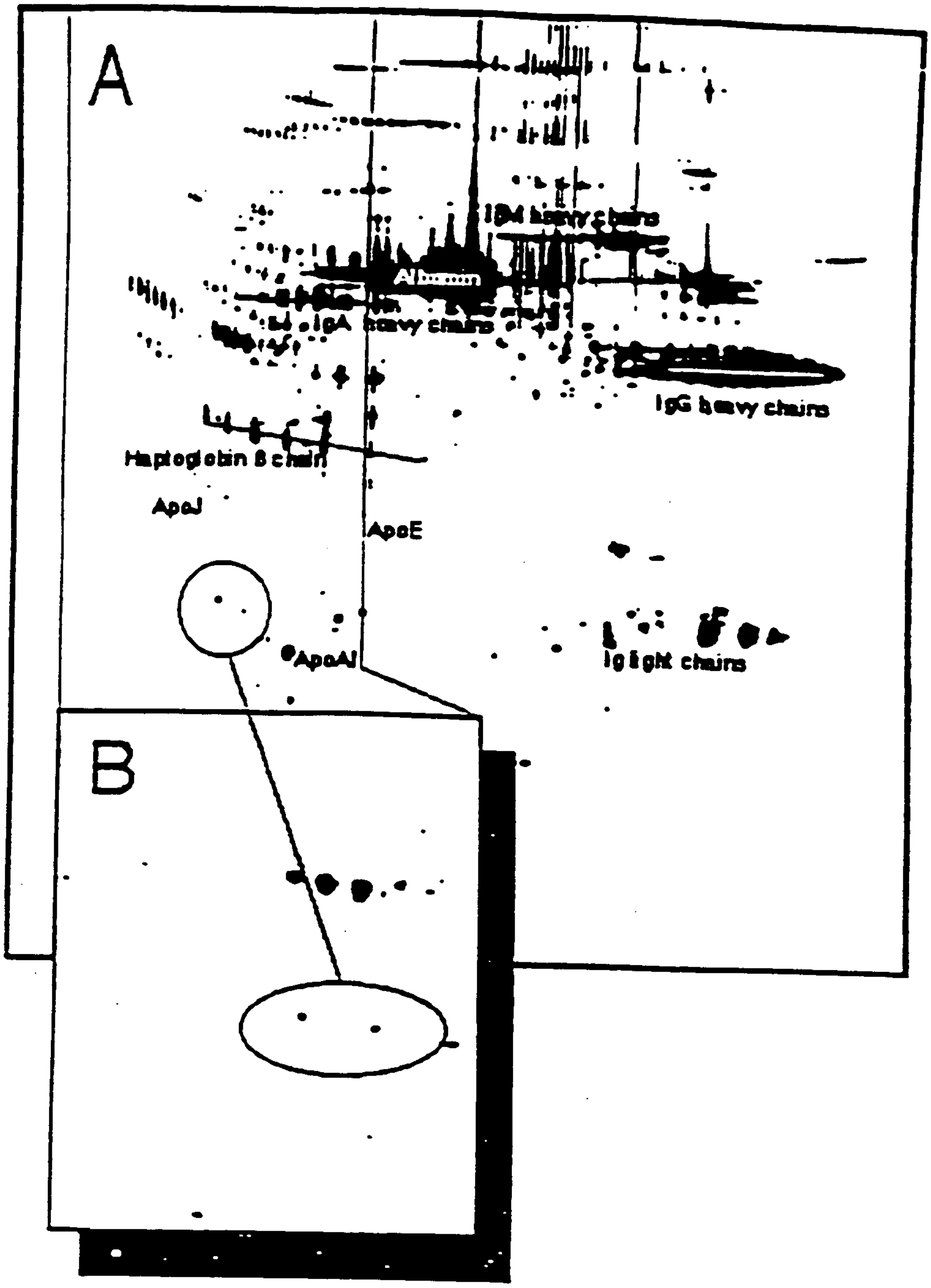


Figure 2. Two-dimensional gel electrophoresis allows 'zooming in' on areas of interest. Rings highlight 2 proteins common to each gel. (A) Wide pI range two dimensional electrophoresis map of human plasma proteins. First dimension separation was achieved using an immobilised pH gradient of 3.5 to 10.0 units. The second dimension was SDS-PAGE. Actual gel size was 16cm x 20cm, and proteins were visualised with silver staining. (B) Narrow pI range electrophoresis was used to 'zoom in' on a small region of the plasma map. The first dimension used a narrow range immobilised pH gradient of 4.2 to 5.2 units, and second dimension was SDS-PAGE. Micropreparative loading was used, and the gel blotted to PVDF. Proteins were visualised with amido black. Actual blot size was 16cm x 20cm

the use of piperazine diacrylyl as a gel crosslinker and the addition of thiosulfate in the catalyst system has been shown to give better resolution and higher sensitivity detection (Hochstrasser and Merril, 1988; Hochstrasser, Patchornik and Merril, 1988).

Notwithstanding the advances described above, there is an increasing demand to improve the reproducibility of 2-D electrophoresis to facilitate database construction and proteome studies. Harrington *et al.* (1993) explain that if a gel resolves 4000 protein spots, and there is 99.5% spot matching from gel to gel, this will produce 20 spot errors per gel. This amount of error, which might accumulate with each gel to gel comparison used in database construction, could produce an unacceptable degree of uncertainty in gel databases. To address these issues, partial automation of large 2-D gel separations has been undertaken (Nokihara, Morita and Kuriki, 1992; Harrington *et al.*, 1993). Although results are preliminary, spot to spot positional reproducibility in one study was found to be threefold improved over manual methods (Harrington *et al.*, 1993). It should be noted that small 2-D gel formats (50 × 43 mm) have been almost completely automated (Brewer *et al.*, 1986), although these are not generally used for database studies.

MICROPREPARATIVE 2-D GEL ELECTROPHORESIS

With the advent of affordable protein microcharacterisation techniques, including N-terminal microsequencing, amino acid analysis, peptide mass fingerprinting, phosphate analysis and monosaccharide compositional analysis, a new challenge for 2-D electrophoresis has been to maintain high resolution and reproducibility but to provide protein in sufficient quantities for chemical analysis (high nanogram to low microgram quantities of proteins per spot). This becomes difficult to achieve with very complex samples such as whole bacterial cells, as the initial protein load is divided among 2000 to 4000 protein species. Two approaches are used for producing amounts of material that can be chemically characterised. The first method is to run multiple gels, collect and pool the spots of interest, and subject them to concentration (Ji *et al.*, 1994; Walsh *et al.*, 1995; Rasmussen *et al.*, 1992). In this approach, the concentration process must also act as a purification step to remove accumulated electrophoretic contaminants such as glycine. A more elegant approach has been to exploit the high loading capacity of IPG isoelectric focusing. The high loading capacity of immobilised pH gradients was described early (Ek, Bjellqvist and Righetti, 1983), but has only recently been applied to 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b). Up to 15 mg of protein can been applied to a single gel, yielding microgram quantities of hundreds of protein species. A further benefit of this approach is that proteins present in low abundance, which may not be visualised by lower protein loads, are more likely to be detected. The use of electrophoretic or chromatographic prefractionation techniques (Hochstrasser *et al.*, 1991a; Harrington *et al.*, 1992), followed by high loading of narrow-range IPG separations (Bjellqvist *et al.*, 1993b) provides a likely solution to studies on proteins present in low abundance.

Methods of protein detection

There are many means for detecting proteins from 2-D gels. The method used will be dictated by factors including protein load on gel (analytical or preparative), the purpose of the gel (for protein quantitation or for blotting and chemical characterisation), and the sensitivity required. The most common means of protein detection and their applications are shown in *Table 1*. Most detection methods have drawbacks, for

Table 1: Common stains for 2-D gels or blots and their applications.

Detection Method	Main applications	Unsuitable applications	Sensitivity	References
[^{35}S] Met or ^{14}C radiolabelling and fluorography or phosphorimaging	Cell lines, cultured organisms	Samples that cannot be labelled	20 ppm of radiolabel in a spot	Garrels and Franza, 1989; Latham, Garrels and Solter, 1993
[^{35}S]thiourea silver	Extremely high sensitivity gel staining	Preparative 2-D, PVDF or NC membranes	0.4 ng protein on spot or band of gel	Wallace and Saluz, 1992a,b
Silver	Very high sensitivity gel staining, can be mono or polychromatic	Preparative 2-D, PVDF or NC membranes	4 ng protein on spot or band of gel	Rabilloud, 1992; Hochstrasser and Merril, 1988
Coomassie blue R-250	Staining of gels; staining of PVDF membranes before protein sequencing	Staining prior to direct mass determination from PVDF; amino acid analysis on PVDF; detection of some glycoproteins	40 ng protein on band or spot of gel	Strupat *et al.*, 1994; Gharahdaghi *et al.*, 1992; Goldberg *et al.*, 1988; Sanchez *et al.*, 1992
Colloidal gold	Staining NC membranes, staining PVDF before direct MALDI-TOF	Gels	60 x higher than coomassie	Yamaguchi and Asakawa, 1988; Eckerskorn *et al.*, 1992; Strupat *et al.*, 1994
Zinc imidazole	Reverse staining of gels or membranes; may be beneficial in MALDI-TOF of peptides	Where positive image is required	Higher than coomassie	Ortiz *et al.*, 1992; James *et al.*, 1993
Ponceau S and amido black	Staining higher protein loads on PVDF, for protein sequencing or amino acid analysis	Staining prior to direct mass determination from PVDF	100 ng protein on band or spot of gel	Sanchez *et al.*, 1992; Strupat *et al.*, 1994; Wilkins *et al.*, 1995
India ink	Staining of membrane-bound proteins, staining PVDF before direct MALDI-TOF	Gel staining, not quantitative from protein to protein	1–10 ng	Li *et al.*, 1989; Hughes, Mack and Hamparian, 1988; Strupat *et al.*, 1994
Stains-all	Staining to detect glycoproteins or Ca^{2+} binding proteins	General gel staining	100 ng protein on band or spot of gel	Campbell, MacLennan and Jorgensen, 1983; Goldberg *et al.*, 1988

PVDF = polyvinylidene difluoride, NC = nitrocellulose, MALDI-TOF = matrix assisted laser desorption ionisation time of flight mass spectrometry.

example, some glycoproteins are not stained by coomassie blue (Goldberg *et al.*, 1988), and many organic dyes are unsuitable for protein detection on PVDF if samples are to be used for direct matrix-assited laser desorption ionisation mass spectrometry (Strupat *et al.*, 1994).

Although most means of protein detection give some indication of the quantities of protein present, in general they cannot be used for global quantitation. This is because

no protein stain is able consistently to detect proteins over a wide range of concentrations, isoelectric points and amino acid compositions, and with a variety of post-translational modifications (Goldberg *et al.*, 1988; Li *et al.*, 1989). Furthermore, there are large differences in staining pattern when identical gels or blots are subjected to different stains, including amido black, imidazole zinc, india ink, ponceau S, colloidal gold, or coomassie blue (Tovey, Ford and Baldo, 1987; Ortiz *et al.*, 1992). The most common means of quantitating large numbers of proteins in a 2-D gel involves the radiolabelling of protein samples prior to electrophoresis, and protein quantitation based on fluorography and image analysis or liquid scintillation counting (Garrels, 1989; Celis and Olsen, 1994). However, proteins which do not contain methionine cannot be detected if only $[^{35}S]$ methionine is used for labelling. Amino acid analysis of protein spots visualised by other techniques presents a likely means of protein quantitation for the future.

BLOTTING OF PROTEINS TO MEMBRANES

Electrophoretic blotting of proteins from two-dimensional polyacrylamide gels to membranes presents many options for protein identification and microcharacterisation which are not possible when proteins remain in gels. For example, when proteins are blotted to polyvinylidene difluoride (PVDF) membranes, they can be identified by N-terminal sequencing, amino acid analysis, or immunoblotting, or they may be subjected to endoproteinase digestion, monosaccharide analysis, phosphate analysis, or direct matrix-assisted laser desorption ionisation mass spectrometry (Matsudaira, 1987; Wilkins *et al.*, 1995; Jungblut *et al.*, 1994; Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Weitzhandler *et al.*, 1993; Murthy and Iqbal, 1991; Eckerskorn *et al.*, 1992). It is possible to combine of some of these procedures on a single protein spot on a PVDF membrane (Packer *et al.*, 1995; Wilkins *et al.*, submitted; Weitzhandler *et al.*, 1993). This is useful when minimal amounts of protein are available for analysis. These techniques will be explored in detail later in this review. Notwithstanding the above, there are some disadvantages associated with blotting of proteins to membranes. There is always loss of sample during blotting procedures (Eckerskorn and Lottspeich, 1993), and common protein detection methods are less sensitive or not applicable to membranes (*Table 1*), presenting difficulties for the analysis of low abundance proteins. Detailed discussion of the merits of available membranes and common blotting techniques can be found elsewhere (Eckerskorn and Lottspeich, 1993; Strupat *et al.*, 1994; Patterson, 1994).

2-D gel analysis, documentation, and proteome databases

Following protein electrophoresis and detection, detailed analysis of gel images is undertaken with computer systems. For proteome projects, the aim of this analysis is to catalogue all spots from the 2-D gel in a qualitative and if possible quantitative manner, so as to define the number of proteins present and their levels of expression. Reference gel images, constructed from one or more gels, form the basis of two-dimensional gel databases. These databases also contain protein spot identities and

details of their post-translational modifications. 2-D gel databases are beginning to be linked to or integrated with comprehensive protein and nucleic acid databases (Neidhardt *et al.*, 1989; Simpson *et al.*, 1992; Appel *et al.*, 1994), and 'organism' databases, containing DNA sequence data, chromosomal map locations, reference 2-D gels and protein functional information for an organism, are becoming established as genome and proteome projects progress (VanBogelen *et al.*, 1992; Yeast Protein Database cited in Garrels *et al.*, 1994).

GEL IMAGE ANALYSIS AND REFERENCE GELS

After 2-D electrophoresis and protein visualisation by staining, fluorography or phosphorimaging, images of gels are digitised for computer analysis by an image scanner, laser densitomer, or charge-coupled device (CCD) camera (Garrels, 1989; Celis *et al.*, 1990a; Urwin and Jackson, 1993). All systems digitise gels with a resolution of 100 – 200 mm, and can detect a wide range of densities or shading (256 or more 'grey scales'). Following this, gel images are subjected to a series of manipulations to remove vertical and horizontal streaking and background haze, to detect spot positions and boundaries, and to calculate spot intensity (*Figure 3*). A standard spot (SSP) number, containing vertical and horizontal positional information, is assigned to each detected spot and becomes the protein's reference number. *Table 2* lists some notable software packages which process 2-D gel images.

Table 2: Some Software Packages for the Analysis of Gel Images.

Gel Image Analysis System	References*
ELSIE 4 & 5	Olsen and Miller, 1988; Wirth *et al.*, 1991; Wirth *et al.*, 1993
GELLAB I & II	Wu, Lemkin and Upton, 1993; Lemkin, Wu and Upton, 1993, Myrick *et al.*, 1993.
MELANIE I & II	Appel, *et al.* 1991, Hochstrasser *et al.* 1991b
QUEST I & II and PDQUEST	Garrels, 1989, Monardo *et al.*, 1994, Holt *et al.*, 1992, Celis *et al.*, 1990a,b
TYCHO & KEPLAR	Anderson *et al.*, 1984, Richardson, Horn and Anderson, 1994

* These references are not exhaustive, they include some references of use as well as authors of the system

As there are difficulties in the electrophoresis of samples with 100% reproducibility, reference gel images are often constructed from many gels of the same sample (Garrels and Franza, 1989; Neidhardt *et al.*, 1989). Since this involves the matching of 2000 to 4000 proteins from one gel to another, it presents a considerable challenge to image analysis systems. Matching of gels is usually initiated by an operator, who manually designates approximately 50 or so prominent spots as 'landmarks' on gels to be cross-matched. Proteins which match are then established around landmarks, using computer-based vector algorithms to extend the matching over the entire gel. Close to 100% of spots from complex samples can be matched by these methods, although different degrees of operator intervention may be required (Olsen and Miller, 1988; Lemkin and Lester, 1989; Garrels, 1989; Myrick *et al.*, 1993).

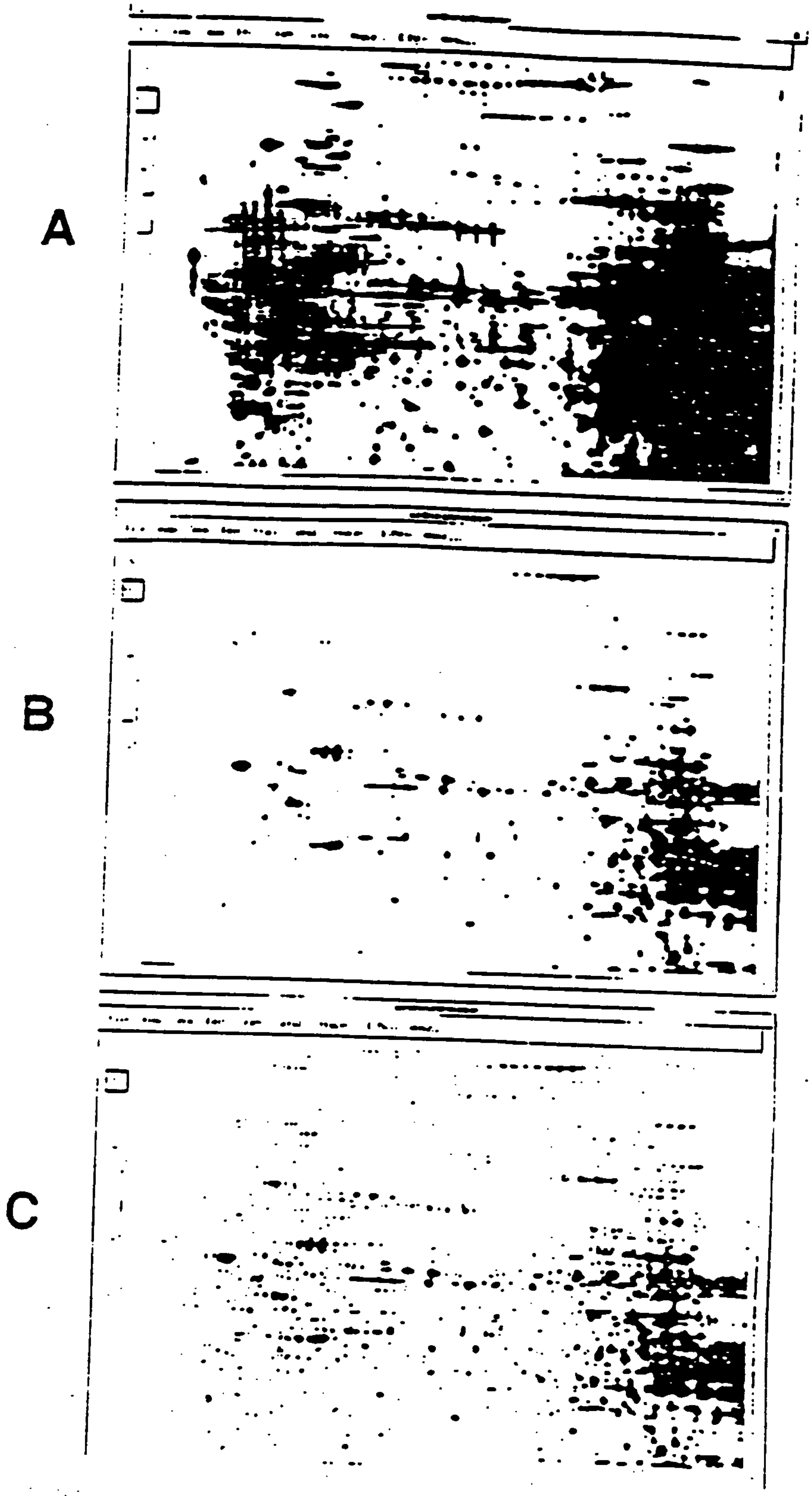


Figure 3. Computer processing of gel images. Shown is a wide pI range 2-D separation of human liver proteins, processed by Melanie software (Appel *et al.*, 1991). (A) Original gel image as captured by laser densitometer. (B) Gel image after processing to remove streaking and background. (C) Outline definition of all spots on the gel.

CALCULATION OF PROTEIN ISOELECTRIC POINT AND MOLECULAR WEIGHT

Estimation of the isoelectric point (pI) and molecular weight (MW) of proteins from 2-D gels provides fundamental parameters for each protein, which are also of use during identification procedures (see following section). The pI and MW of proteins are recorded in 2-D gel databases. Accurate estimations of protein pI and MW can be obtained by using 20 or more known proteins on a reference map to construct standard curves of pI and molecular weight, which are then used to calculate estimated pI and MW of unknown proteins (Neidhardt *et al.*, 1989; Garrels and Franza, 1989; VanBogelen, Hutton and Neidhardt, 1990; Anderson and Anderson, 1991; Anderson *et al.*, 1991; Latham *et al.*, 1992). Alternatively, the MW of individual proteins blotted to PVDF can be determined very accurately by direct mass spectrometry (Eckerskorn *et al.*, 1992). Where immobilised pH gradients are used, the focusing position of proteins allows their pI to be measured within 0.15 units of that calculated from the amino acid sequence (Bjellqvist *et al.*, 1993c). It must be noted, however, that proteins carrying post-translational modifications may migrate to unexpected pI or MW positions during electrophoresis (Packer *et al.*, 1995).

SPOT QUANTITATION AND EXPRESSION ANALYSIS

A major challenge faced in proteome projects is the quantitative analysis of proteins separated by 2-D electrophoresis. The most accurate means of protein quantitation is to determine chemically the amount of each protein present by amino acid compositional analysis. However, the current method of choice for quantitative analysis of many proteins is to radiolabel samples with [^{35}S] methionine or ^{14}C amino acids, perform the 2-D electrophoresis, and measure protein levels in disintegrations per minute (dpm) or units of optical density. Quantitation is achieved either by liquid scintillation counting, or by gel image analysis where spot densities are quantitated by reference to gel calibration strips containing known amounts of radiolabelled protein or against the integrated optical density of all spots visualised (Vandekerkhove *et al.*, 1990; Celis *et al.*, 1990b; Celis and Olsen, 1994; Garrels, 1989; Latham, Garrels and Solter, 1993; Fey *et al.*, 1994). All approaches effectively allow spots to be normalised against the total disintegrations per minute loaded onto the gel. Limitations that remain with radiolabelling methods are that absolute quantitation is not achieved because all proteins have varying amounts of any amino acid, and that only easily labelled samples can be investigated. Quantitative silver staining presents an alternative (Giometti *et al.*, 1991; Harrington *et al.*, 1992, Rodriguez *et al.*, 1993; Myrick *et al.*, 1993), which when undertaken with [^{35}S]thiourea (Wallace and Saluz, 1992 a,b) is of extremely high sensitivity.

When protein spots from samples prepared under different conditions are quantitated and matched from gel to gel, it becomes possible to examine changes and patterns in protein expression. Large scale investigation of up- and down-regulation of proteins, their appearance and disappearance, can be undertaken. For example, simian virus 40 transformed human keratinocytes were shown to have 177 up-regulated and 58 down-regulated proteins compared to normal keratinocytes (Celis and Olsen, 1994); detailed synthesis profiles of 1200 proteins have been established in 1 to 4 cell mouse embryos (Latham *et al.*, 1991, 1992); and 4 proteins out of 1971 were found to be markers for

cadmium toxicity in urinary proteins (Myrick *et al*., 1993). Complex global changes in protein expression as a result of gene disruptions have also been investigated (S. Fey and P. Mose-Larsen, Personal communication). Impressively, large gel sets showing protein expression under different conditions can be globally investigated using statistical methods that find groups of related objects within a set. For example, the REF52 rat cell line database, consisting of 79 gels from 12 experimental groups where each gel contains quantitative data for 1600 cross-matched proteins, has been analysed by cluster analysis (Garrels *et al*., 1990). This revealed clusters of proteins that, for example, were induced or repressed similarly under simian virus 40 or adenovirus transformation, suggesting a common mechanism. Protein groups that were induced or repressed during culture growth to confluence were also found. It is obvious that the potential for investigation of cellular control mechanisms by these approaches is immense. It is equally clear that investigations of gene expression of this scale are currently technically impossible using nucleic-acid based techniques.

Table 3: Some proteome databases and their special features

Proteome database	Special features	References
E. coli gene-protein database	Gel spots linked with GenBank and Kohara clones; quantitative spot measurements under different growth conditions	VanBogelen and Neidhardt, 1991, VanBogelen *et al*., 1992
Human heart databases	Identification of disease markers; two separate databases have been established	Baker *et al*., 1992 Corbett *et al*., 1994b Jungblut *et al*., 1994
Human keratinocyte database	Extensive identifications; quantitative spot measurements of transformed cells; identification of disease markers	Celis *et al*., 1990a Celis *et al*., 1993 Celis and Olsen 1994
Mouse embryo database	Quantitative spot measurements through 1 to 4 cell stage	Latham *et al*., 1991 Latham *et al*., 1992
Mouse liver database (Argonne Protein Mapping Group)	Documents changes due to exposure to ionizing radiation and toxic chemicals	Giometti, Taylor and Tollaksen, 1992
Rat liver epithelial database	Detailed subcellular fractionation studies	Wirth *et al*., 1991 Wirth *et al*., 1993
Rat liver database	Extensive studies on regulation of proteins by drugs and toxic agents	Anderson and Anderson, 1991, Anderson *et al*., 1992; Richardson, Horn and Anderson, 1994
REF 52 rat cell line database	Accessible via World Wide Web, quantitative spot measurements under different conditions	Garrels and Franza 1989 Boutell *et al*., 1994
SWISS-2DPAGE containing human reference maps	Accessible via World Wide Web, completely integrated with SWISS-PROT and SWISS-3DIMAGE	Appel *et al*., 1993 Hochstrasser *et al*., 1992 Hughes *et al*., 1993 Golaz *et al*., 1993
Yeast Protein Database (YPD) and Yeast Electrophoretic Protein Database (YEPD)	Completely crossreferenced organism database; YPD has extensive information on over 3500 proteins; YEPD has many identifications	Garrels *et al*., 1994

FEATURES OF PROTEOME DATABASES

Proteome projects rely heavily on computer databases to store information about all proteins expressed by an organism. 'Proteome databases' should contain detailed information of proteins already characterised elsewhere, as well as protein data from 2-D gels such as apparent pI and MW, expression level under different conditions, subcellular localisation, and information on post-translational modifications. Images of reference 2-D gels, showing protein SSP numbers and protein identifications, should also be included. Ideally, proteome databases should be accessible with Macintosh or IBM personal computers and easy to use. Some proteome databases and the areas they cover are listed in *Table 3*. Databases range from collections of annotated gels to large databases of images integrated with protein and nucleic acid sequence banks.

One example of an integrated proteome database is the suite of SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE databases (Appel *et al.*, 1993; Appel *et al.*, 1994; Appel, Bairoch and Hochstrasser, 1994; Bairoch and Boeckmann, 1994). The features of these three databases are listed in *Table 4*. SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE are accessible through the World Wide Web

Table 4: The SWISS-PROT, SWISS-2DPAGE and SWISS-3DIMAGE suite of crosslinked databases. All three databases are accessible through the World Wide Web, at URL address: http://expasy.hcuge.ch/

	SWISS-PROT	SWISS-2DPAGE	SWISS-3DIMAGE
Information	Text entries of sequence data; Citation information; taxonomic data. 38, 303 entries in Release 29	2-D gel images of: human liver, plasma, HepG2, HepG2 secreted proteins, red blood cell, lymphoma, cerebrospinal fluid, macrophage like cell line, erythroleukemia cell, platelet	Collection of 330 3-D images of proteins
Annotations	Protein function. Post translational modifications. Domains; Secondary structure. Quaternary structure. Diseases associated with protein. Sequence conflicts	Gel images where protein is found. How protein identified. Protein pI and MW, protein number; normal and pathological variants	All annotation is available in SWISS-PROT
Cross-Referenced Databases	SWISS-2DPAGE SWISS-3DIMAGE EMBL, PIR, PDB, OMIM, PROSITE, Medline, Flybase; GCRDb, MaizeDB, WormPep, DictyDB	SWISS-PROT and all other databases accessible through SWISS-PROT	SWISS-PROT and all other databases accessible through SWISS-PROT
Other Features	Navigation to other SWISS databases achieved by selecting entries with computer mouse	Gel images show position of identified proteins, or region of gel where protein should appear	Mono and stereo images available. Images can be transferred to local computer image viewing programs

(Berners-Lee *et al.*, 1992), allowing any computer connected to the internet to access the stored information and images. Navigation within and between the three databases is seamless, as all potential crosslinks are highlighted as hypertext on the display and can be selected with a computer mouse. From these databases, detailed information about a protein, including amino acid sequence and known post-translational modifications, can be obtained, the precise protein spot it corresponds to on a reference gel image can be viewed if known, and the 3-D structure of the molecule can be seen if available. References to nucleic acid and other databases are also given to provide access to information stored elsewhere.

Organism' databases, containing detailed protein and nucleic acid information about a species, are becoming common as genome and proteome projects progress. These differ from nucleic acid or protein sequence databases like GenBank or SWISS-PROT because they are image based, and contain information about chromosomal map positions, transcription of genes, and protein expression patterns. The *Escherichia coli* gene-protein database (VanBogelen, Hutton and Neidhardt, 1990; VanBogelen and Neidhardt, 1991, VanBogelen *et al.*, 1992), known as the ECO2DBASE, is one example. It contains gene and protein names, 2-D gel spot information (including pI and MW estimates, and spot identification), genetic information (GenBank or EMBL codes, chromosomal location, location on Kohara clones (Kohara, Akiyama, and Isono, 1987), transcription direction of genes), and protein regulatory information (level of protein expression under different growth regimes, member of regulon or stimulon). All entries in the ECO2DBASE are also cross-referenced to the SWISS-PROT database (Bairoch and Boeckmann, 1994). It is anticipated that organism databases will soon become a standard means of storing all available information about a particular species. However there is currently no consistent manner in which organism databases are assembled, which may hamper comparisons in the future.

Identification and characterisation of proteins from 2-D gels

The number of proteins identified on a 2-D reference map determines its usefulness as a research and reference tool. As most reference maps have only a small proportion of proteins identified, a major aim of current proteome projects is to screen many proteins from 2-D maps, in order to define them as 'known' in current nucleic acid and protein databases, or as 'unknown'. Protein identification assists in confirmation of DNA open reading frames, and provides focus for DNA sequencing projects and protein characterisation efforts by pointing to proteins that are novel. Since there may be 3000–4000 proteins from a single 2-D map that require identification, the challenge in protein screening is to identify proteins quickly, with a minimum of cost and effort.

Traditionally, proteins from 2-D gels have been identified by techniques such as immunoblotting, N-terminal microsequencing, internal peptide sequencing, comigration of unknown proteins with known proteins, or by overexpression of homologous genes of interest in the organism under study (Matsudaira, 1987; Rosenfeld *et al.*, 1992; VanBogelen *et al.*, 1992; Celis *et al.*, 1993; Honore *et al.*, 1993; Garrels *et al.*, 1994). Whilst these techniques are powerful identification tools, they are too expensive or time and labour intensive to use in mass screening programs. A hierarchical approach to mass protein identification has been recently suggested as an

Table 5: Hierarchical analysis for mass screening of 2-D separated proteins blotted to membranes. Rapid and inexpensive techniques are used as a first step in protein identification, and slower more expensive techniques are then used if necessary. Table modified from Wasinger *et al.*, 1995.

Order	Identification technique	References
1	Amino acid analysis	Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Jungblut *et al.*, 1994; Wilkins *et al.*, 1995
2	Amino acid analysis with N-terminal sequence tag	Wilkins *et al.*, submitted
3	Peptide-mass fingerprinting	Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995
4	Combination of amino acid analysis and peptide mass fingerprinting	Cordwell *et al.*, 1995; Wasinger *et al.*, 1995
5	Mass spectrometry sequence tag	Mann and Wilm, 1994
6	Extensive N-terminal Edman microsequencing	Matsudaira, 1987
7	Internal peptide Edman microsequencing	Rosenfeld *et al.*, 1992; Hellman *et al.*, 1995
8	Microsequencing by mass spectrometry (electrospray ionisation, post-source decay MALDI-TOF)	Johnson and Walsh, 1992
9	Ladder sequencing	Bartlet-Jones *et al.*, 1994

alternative to traditional approaches (*Table 5*; Wasinger *et al.*, 1995). This involves the use of rapid and cheap identification tools such as amino acid analysis and peptide mass fingerprinting as first steps in protein identification, followed by the use of slower, more expensive and time consuming identification procedures if necessary. In the construction of this hierarchy the analysis time, cost per sample and the complexity of the data created has been considered, as whilst some techniques require little machine time per sample, the analysis of data can be quite involved and time consuming. Amino acid analysis and peptide mass-fingerprinting based identification techniques in the hierarchy are discussed in detail below. For review of other protein identification techniques in *Table 5*, see Patterson (1994) and Mann (1995).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION

There has been a revival of interest in the use of amino acid composition for identification of proteins from 2-D gels after early work by Eckerskorn *et al.* (1988). This technique uses a protein's idiosyncratic amino acid composition profile in order to identify it by comparison with theoretical compositions of proteins in databases. The amino acid composition of proteins can be determined by differential metabolic radiolabelling and quantitative autoradiography after 2-D electrophoresis (Garrels *et al.*, 1994; Frey *et al.*, 1994), or by acid hydrolysis of membrane-blotted proteins and chromatographic analysis of the resulting amino acid mixture (Eckerskorn *et al.*, 1988; Tous *et al.*, 1989; Gharahdaghi *et al.*, 1992; Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). As differential metabolic labelling experiments require X-ray film or phosphor-image plate exposures of up to 140 days, and can only be undertaken with easily radiolabelled samples, the technique is not as rapid or widely applicable as chromato-

```
Spot: ECOLI-E1M
==============

Composition:

Asx: 13.2   Glx: 10.4   Ser:  5.7   His:  0.7
Gly:  5.4   Thr:  3.6   Ala:  6.7   Pro:  7.9
Tyr:  1.3   Arg:  5.0   Val:  8.0   Met:  0.3
Ile:  5.9   Leu:  8.0   Phe: 13.3   Lys:  4.4

pI estimate:   6.89  Range searched: ( 6.64,  7.14)
Mw estimate:  16800  Range searched: (13640, 20160)

Closest SWISS-PROT entries for the species ECOLI matched by AA composition:
```

Rank	Score	Protein	pI	Mw	Description
1	**24**	**PYRI_ECOLI**	**6.84**	**16989**	**ASPARTATE CARBAMOYLTRANSFERASE**
2	39	COAA_ECOLI	6.32	36359	PANTOTHENATE KINASE (EC 2.7.1.33)
3	40	META_ECOLI	5.06	35713	HOMOSERINE O-SUCCINYLTRANSFERASE
4	42	CADC_ECOLI	5.52	57812	TRANSCRIPTIONAL ACTIVATOR CADC.
5	43	HLYC_ECOLI	8.56	19769	HEMOLYSIN C. PLASMID.

```
Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:
```

Rank	Score	Protein	pI	Mw	Description
1	**24**	**PYRI_ECOLI**	**6.84**	**16989**	**ASPARTATE CARBAMOYLTRANSFERASE**
2	102	TRAJ_ECOLI	6.73	17921	TRAJ PROTEIN.
3	112	YAJG_ECOLI	6.79	19028	HYPOTHETICAL LIPOPROTEIN YAJG.
4	140	YFJB_ECOLI	6.83	14945	HYPOTHETICAL 14.9 KD PROTEIN IN GRPE
5	142	YAHA ECOLI	7.06	14726	HYPOTHETICAL PROTEIN IN BETT 3'REGION

Figure 4. Computer printout from ExPASy server where the empirical amino acid composition, estimated pI and MW of a protein from a 2-D reference map of *E. coli* were matched against all entries in SWISS-PROT for *E. coli*. The correct identification, aspartate carbamoyltransferase, is shown in bold. Low scores indicate a good match. Note how matching within a defined pI and MW range (lower set of proteins) has greatly increased the score difference between the first and second ranking proteins. This score difference gives high confidence in the identification, and is only observed where the top ranking protein is the correct identification (Wilkins *et al.*, 1995).

graphy-based analysis. Proteins blotted to PVDF membranes can be hydrolysed in 1 h at 155°C, amino acids extracted in a single brief step, and each sample automatically derivatised and separated by chromatography in under 40 minutes (Wilkins *et al.*, 1995; Ou *et al.*, 1995). In this manner, one operator can routinely analyse 100 proteins per week on one HPLC unit. This technology lends itself to automation, and it is anticipated that instruments with even greater sample throughput will be developed. When proteins have been prepared by micropreparative 2-D electrophoresis (Hanash *et al.*, 1991; Bjellqvist *et al.*, 1993b), blotted to a PVDF membrane and stained with amido black, any visible protein spot is of sufficient quantity for amino acid analysis (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995).

After the amino acid composition of a protein has been determined, computer programs are used to match it against the calculated compositions of proteins in databases (Eckerskorn *et al.*, 1988; Sibbald, Sommerfeldt and Argos, 1991; Jungblut *et al.*, 1992; Shaw, 1993; Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995). Matching is usually done with only 15 or 16 amino acids, as cysteine and

```
Spot ECOLI-ACJ
==============

Composition:

Asx:  9.6   Glx: 10.8   Ser:  4.1   His:  2.7
Gly: 11.2   Thr:  2.8   Ala: 11.9   Pro:  3.2
Tyr:  6.0   Arg:  3.7   Val:  9.5   Met:  0.6
Ile:  5.6   Leu:  8.2   Phe:  3.2   Lys:  4.9

pI estimate:   5.99  Range searched: ( 5.74,  6.24)
Mw estimate:  45000  Range searched: (36000, 54000)

Closest SWISS-PROT entries for ECOLI with pI and Mw values in specified
range:
```

Rank	Score	Protein	pI	Mw	N-terminal Seq.
1	21	GLYA_ECOLI	6.03	45316	M L K R E
2	32	YJGB_ECOLI	5.86	36502	M S M I K
3	38	GABT_ECOLI	5.78	45774	M S N S K
4	44	YIMS_ECOLI	5.86	48018	M R I K Y
5	45	DHE4_ECOLI	5.98	48581	M D Q T Y
6	46	ARGD_ECOLI	5.79	43765	M A I E Q
7	46	MJPB_ECOLI	5.78	37851	M N H S L
8	47	GLMU_ECOLI	5.98	49162	M L N N A
9	47	ACKA_ECOLI	5.85	43290	M S S K L
10	50	YJJN_ECOLI	6.01	37064	M E S K I

Figure 5. A PVDF protein spot from an *E. coli* 2-D reference map was sequenced for 4 cycles, and the same sample then subject to amino acid analysis. The N-terminal sequence was M L K R. When the amino acid composition of the spot, as well as estimated pI and MW, were matched against all entries in SWISS-PROT for *E. coli*, the above list of best matches was produced. N-terminal sequences are from SWISS-PROT for those entries. The top ranking identification of serine hydroxymethyltransferase (bold) did not show a large score difference between the first and second ranking proteins, giving little confidence in this being the correct protein identification. However, the sequence tag (M L K R) confirmed the identity of the protein as serine hydroxymethyltransferase.

tryptophan are destroyed during hydrolysis, asparagine and glutamine are deamidated to their corresponding acids, and proline is not quantitated in some analysis systems. The computer programs produce a list of best matching proteins, which are ranked by a score that indicates the match quality. Some programs allow matching to be restricted to specific 'windows' of MW and pI (Hobohm, Houthaeve and Sander, 1994; Wilkins *et al.*, 1995), and to protein database entries for one species (Jungblut *et al.*, 1992; Wilkins *et al.*, 1995). The use of such restrictions increases the power of matching. An example of protein identification by amino acid composition is shown in *Figure 4*. To date, amino acid composition has been used to identify proteins from reference maps of *Spiroplasma melliferum*, *Mycoplasma genitalium*, *E. coli*, *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, human sera, human heart, human lymphocyte, and mouse brain (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Wilkins *et al.*, 1995; Jungblut *et al.*, 1992, 1994; Garrels *et al.*, 1994; Frey *et al.*, 1994).

PROTEIN IDENTIFICATION BY AMINO ACID COMPOSITION AND N-TERMINAL SEQUENCE TAG

When samples from 2-D gels are not unambiguously identified by amino acid

composition, pI and MW, often the correct identification of that protein is amongst the top rankings of the list (Hobohm, Houthaeve and Sander, 1994; Cordwell *et al.*, 1995; Wilkins *et al.*, 1995). Taking advantage of this observation, we have used the mass spectrometry 'sequence tag' concept (Mann and Wilm, 1994) in developing a combined Edman degradation and amino acid analysis approach to protein identification (Wilkins *et al.*, submitted). This involves the N-terminal sequencing of PVDF-blotted proteins by Edman degradation for 3 or 4 cycles to create a 'sequence tag', following which the same sample is used for amino acid analysis. As only a few amino acids are removed from the protein, its composition is not significantly altered. Furthermore, since only a small amount of protein sequence is required, fast but low repetitive yield Edman degradation cycles can be used. Modifications to current procedures should allow 3 cycles to be completed in 1 h, thereby allowing the screening of 100 or more proteins per week on one automated, multi-cartridge sequenator. Amino acid composition, pI and MW of proteins are matched against databases as described above, and N-terminal sequences of best matching proteins are checked with the 'sequence tag' to confirm the protein identity (*Figure 5*). This technique will be less useful when proteins are N-terminally blocked, but as only a few N-terminal amino acids are susceptible to the acetyl, formyl, or pyroglutamyl modifications that cause blockage, this may itself provide useful information for sequence tag identification. A strength of N-terminal sequence tag and amino acid composition protein identification is that data generated are quickly and easily interpreted.

PROTEIN IDENTIFICATION BY PEPTIDE MASS FINGERPRINTING

Techniques for the identification of proteins by peptide mass fingerprinting have recently been described (Henzel *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; James *et al.*, 1993; Mann, Hojrup and Roepstorff, 1993; Yates *et al.*, 1993; Mortz *et al.*, 1994; Sutton *et al.*, 1995). This involves the generation of peptides from proteins using residue-specific enzymes, the determination of peptide masses, and the matching of these masses against theoretical peptide libraries generated from protein sequence databases. As proteins have different amino acid sequences, their peptides should produce characteristic 'fingerprints'.

The first step of peptide mass fingerprinting is protein digestion. Proteins within the gel matrix or bound to PVDF can be enzymatically digested *in situ*, although *in situ* gel digests are reported to produce more enzyme autodigestion products, which complicate subsequent peptide mass analysis (James *et al.*, 1993; Rasmussen *et al.*, 1994; Mortz *et al.*, 1994). The enzyme of choice for digestion is currently trypsin (of modified sequencing grade), but other enzymes (Lys-C or *S. aureus* V8 protease) have also been used (Pappin, Hojrup and Bleasby, 1993). To maximise the number of peptides obtained, it is desirable for protein samples to be reduced and alkylated prior to digestion (Mortz *et al.*, 1994; Henzel *et al.*, 1993). This ensures that all disulfide bonds of the protein are broken, and produces protein conformations that are more amenable to digestion. Surprisingly, chemical digestion methods such as cyanogen bromide (methionine specific), formic acid (aspartic acid specific), and 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine (tryptophan specific) have not been explored as means of peptide production for mass fingerprinting, even though they are rapid and may circumvent some problems associated with enzyme digestions

(Nikodem and Fresco, 1979; Crimmins *et al.*, 1990; Vanfleteren *et al.*, 1992).

After proteins are digested, peptide masses are determined by mass spectrometry. Direct analysis of peptide mixtures can be achieved by electrospray ionisation mass spectrometry, plasma desorption mass spectrometry, or matrix assisted laser desorption ionization (MALDI) mass spectrometry techniques. MALDI is preferable because of its higher sensitivity and greater tolerance to contaminating substances from 2-D gels (James *et al.*, 1993; Mortz *et al.*, 1994; Pappin, Hojrup and Bleasby, 1993). Furthermore, recent modifications to sample preparation methods have largely solved early difficulties experienced with the calibration of MALDI spectra (Mortz *et al.*, 1994; Vorm and Mann, 1994; Vorm, Roepstorff and Mann, 1994). The high sensitivity of mass spectrometry allows a small fraction of a digest of a 1μg protein spot to be used for analysis, and analysis itself is complete in a few minutes.

A major challenge associated with peptide mass fingerprinting is data interpretation prior to computer matching against libraries of theoretical peptide digests. Spectra must be examined carefully to determine which peaks represent peptide masses of interest, as there are often enzyme autodigestion products and contaminating substances present (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). Furthermore, if protein alkylation and reduction has not been undertaken prior to protein digestion, peptide sequence coverage may be poor (40% to 70%), with some masses present representing disulfide bonded peptides originally present in the protein (Mortz *et al.*, 1994). For eukaryotes, a serious issue is the alteration of peptide masses by the presence of post-translational modifications (*Table 6*). The mass of the unmodified peptide alone can be very difficult to determine. Two artifactual modifications introduced by electrophoresis, an acrylamide adduct to cysteine and the oxidation of methionine, are also known to alter peptide masses (le Maire *et al.*, 1993; Hess *et al.*, 1993).

Table 6: Masses of some common post-translational modifications. Peptides carrying post-translational modifications complicate data analysis for peptide mass fingerprinting protein identification. This is especially so for protein glycosylation, which involves many different combinations of the hexosamines, hexoses, deoxyhexoses, and sialic acid

Post-translational modification	Mass change
Acetylation	+42.04
*Acrylamide adduct to cysteine	+71.00
Carboxylation of Asp or Glu	+44.01
Deamidation of Asn or Gln	+0.98
Disulfide bond formation	-2.02
Deoxyhexoses (Fuc)	146.14
Formylation	+28.01
Hexosamines (GlcN, GalN)	+161.16
Hexoses (Glc, Gal, Man)	+162.14
Hydroxylation	+16.00
N-acetylhexosamines (GlcNAc, GalNAc)	+203.19
*Oxidation of Met	+16.00
Phosphorylation	+79.98
Pyroglutamic acid formed from Gln	-17.03
Sialic acid (NeuNAc)	+291.26
Sulfation	+80.06

Table modified from Finnigan LASERMAT application data sheet 5.
Asterisk * shows modifications that can arise artifactually from the 2-D electrophoresis process.

A number of computer programs are available for matching peptide masses against databases (reviewed in Cottrell, 1994). Matching is usually undertaken in an interactive manner, whereby peaks of mass 500–3000 Da are selected and matched under various search parameters including MW of protein, mass accuracy of peptides, and number of missed enzyme cleavages allowed (Henzel *et al.*, 1993; Mortz *et al.*, 1994; Rasmussen *et al.*, 1994). The correct protein identity is the protein which has the most peptide masses in common with the unknown sample. Identities have been established with as few as three peptides, but unambiguous identification is thought to require a mass spectrometric map covering most peptides of the protein (Mortz *et al.*, 1994; Yates *et al.*, 1993). To date, peptide mass fingerprinting of proteins has been undertaken from the human myocardial protein and keratinocyte maps, from an *E. coli* 2-D gel, and from reference maps of *Spiroplasma melliferum* and *Mycoplasma genitalium* (Sutton *et al.*, 1995; Rasmussen *et al.*, 1994; Henzel *et al.*, 1993; Cordwell *et al.*, 1995, Wasinger *et al.*, 1995), although the technique is most powerful when used in combination with another protein identification technique (Rasmussen *et al.*, 1994; Cordwell *et al.*, 1995).

MASS SPECTROMETRY SEQUENCE TAGGING

An extension of peptide mass fingerprinting has recently been described, called peptide sequence tagging (Mann and Wilm, 1994; Mann, 1995). This uses tandem mass spectrometry (MS/MS) to initially determine the mass of peptides, then subject them to fragmentation by collision with a gas, and finally determine the mass of fragments. The resulting spectra gives information about a peptide's amino acid sequence. The fragmentation masses of peptides can rarely be used to assign a complete sequence, but it usually allows a short 'sequence tag' of 2 or 3 amino acids to be determined. This sequence tag and the original peptide mass is matched by computer against a database, providing a likely identity of the peptide and the protein it came from. The major drawback for this technique as a mass screening tool is the complexity of the mass data generated and the high level of expertise required for its interpretation. Nevertheless, it represents a useful new protein identification method which greatly increases the power of peptide mass fingerprinting protein identification.

Cross-species protein identification

Protein sequence databases continue to grow at a rapid rate, yet it is not widely appreciated that close to 90% of all information contained in current protein databases comes from only 10 species (A. Bairoch, Pers. Comm.). Fortunately, this information can be used to study proteomes of organisms that are poorly defined at the molecular level, via 2-D electrophoresis and 'cross-species' protein identification (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995). This approach allows proteins from reference maps of many different species to be identified without the need for the corresponding genes to be cloned and sequenced. This is particularly true for 'housekeeping' proteins, such as enzymes involved in glycolysis, DNA manipulation and protein manufacture, which are highly conserved across species boundaries. Proteins that cannot be identified across species boundaries can then become the focus of further protein characterisation and DNA sequencing efforts.

A)

Protein APA1_HUMAN
==================

Asx:	8.4	Glx:	19.3	Ser:	6.3	His:	1.3
Gly:	4.2	Thr:	4.3	Ala:	8.0	Pro:	4.2
Tyr:	2.9	Arg:	6.7	Val:	5.5	Met:	1.3
Ile:	0.0	Leu:	15.5	Phe:	2.5	Lys:	8.8

pI Range: no range specified
Mw Range: no range specified

The closest SWISS-PROT entries are:

Rank	Score	Protein	(pI	Mw)	Description
1	0	APA1_HUMAN	5.27	28078	APOLIPOPROTEIN A-I.
2	4	APA1_MACFA	5.43	28005	APOLIPOPROTEIN A-I.
3	12	APA1_RABIT	5.25	27836	APOLIPOPROTEIN A-I.
4	14	APA1_BOVIN	5.36	27549	APOLIPOPROTEIN A-I.
5	14	APA1_CANFA	5.10	27467	APOLIPOPROTEIN A-I.
6	18	APA1_MOUSE	5.42	27922	APOLIPOPROTEIN A-I.
7	26	APA1_PIG	5.19	27598	APOLIPOPROTEIN A-I.
8	27	APA1_CHICK	5.26	27966	APOLIPOPROTEIN A-I.
9	37	DYNA_CHICK	5.44	117742	DYNACTIN, 117 KD ISOFORM.
10	39	APA4_HUMAN	5.18	43374	APOLIPOPROTEIN A-IV.

B)

Reagent: Trypsin MW filter: 10%

Scan using fragment mws of:

1953	1933	1731	1613	1401	1387
1301	1283	1252	1235	1231	1215
1031	996	873	831	813	781
732	704				

No. of database entries scanned = 72018

1	APA1_HUMAN	APOLIPOPROTEIN A-I (APO-AI). - HOMO SAPIENS
2	APA1_MACFA	APOLIPOPROTEIN A-I (APO-AI). - MACACA FASCICULARIS
3	APA1_PAPHA	APOLIPOPROTEIN A-I (APO-AI). - PAPIO HAMADRYAS
4	B41845	orf B - Treponema denticola
5	APA1_CANFA	APOLIPOPROTEIN A-I (APO-AI). - CANIS FAMILIARIS (DOG).
6	S30947	hypothetical protein 1 - Azotobacter vinelandii
7	HS22_PEA	CHLOROPLAST HEAT SHOCK PROTEIN PRECURSOR. - PISUM SATIVU
8	S20724	Tropomyosin - African clawed frog
9	HIVVI354	HIVVI354 premature term. at 793 - Human immunodeficiency
10	TRAJ_ECOLI	TRAJ PROTEIN. - ESCHERICHIA COLI.

Figure 6. Theoretical cross-species matching of human apolipoprotein A-I by amino acid composition and tryptic peptides. When an unknown protein is analysed, best ranking proteins from both techniques can be compared. If the same protein type is observed in both lists, there is high confidence in this being the identity of the unknown molecule (Cordwell *et al.*, 1995). (A) Output of ExPASy server (Appel, Bairoch and Hochstrasser, 1994) where the true amino acid composition of apolipoprotein A-I was matched against all entries in the SWISS-PROT database, without pI or MW windows. Seven of the top 10 matching proteins were apolipoprotein A-I of different species. (B) Output of MOWSE peptide mass fingerprinting program (Pappin, Hojrup and Bleasby, 1993) where true tryptic peptides of human apolipoprotein A-I were matched against the OWL database, using MW window of 10%. Four of the top ten matching proteins were apolipoprotein A-I from different species.

Rapid cross-species identification of proteins from 2-D reference maps can be undertaken with amino acid composition or peptide mass fingerprinting methods (*Figure 6*), but these techniques alone may not identify proteins unambiguously when phylogenetic cross-species distances are great or analysis data is of poor quality (Yates *et al*., 1993; Shaw, 1993; Cordwell *et al*., 1995). However, very high confidence in protein identities can be achieved when lists of best-matching proteins generated by both techniques are compared (Cordwell *et al*., 1995; Wasinger *et al*., 1995). The correct identification is found when the same protein is ranked highly in lists of best matches generated by both techniques. This method has allowed approximately 120 proteins from the reference map of the mollicute *Spiroplasma melliferum*, representing approximately one quarter of the proteome, to be confidently identified by reference to protein information from other species (S. Cordwell, Personal Communication). When cross-species protein identification is to be undertaken, it should be noted that the molecular weight of a protein type across species is usually highly conserved, but that protein pI can vary by more than 2 units (Cordwell *et al*., 1995). Accurate molecular weight determination by direct mass spectrometry of proteins blotted to PVDF (Eckerskorn *et al*., 1992) should therefore be a useful additional parameter for cross-species protein identification.

CHARACTERISATION OF POST-TRANSLATIONAL MODIFICATIONS

Many proteins are modified after translation. Such post-translational modifications, including glycosylation, phosphorylation, and sulfation (see *Table 6*), are usually necessary for protein function or stability. Some abnormal modifications are associated with disease (Duthel and Revol, 1993; Ghosh *et al*., 1993; Yamashita *et al*., 1993). In proteome studies, post-translational modifications can be examined on all proteins present, or on individual spots. Studies on all proteins provide an indication of which proteins may carry a certain type of modification. For example, 2-D gel analysis of cell cultures grown in the presence of [^{3}H] mannose or [^{32}P] phosphate gives an indication of which proteins carry glycans containing mannose, and which proteins are phosphorylated (Garrels and Franza, 1989). Lectin binding studies of 2-D gels blotted to PVDF or nitrocellulose provide information on the saccharides, if any, that are carried by proteins present (Gravel *et al*., 1994).

When individual proteins of interest carrying post-translational modifications have been found, micropreparative 2-D electrophoresis can be used to purify them in microgram quantities (Hanash *et al*., 1991; Bjellqvist *et al*., 1993b). If protein isoforms of similar MW and pI are to be studied, focusing with narrow range pI gradients (1 pH unit) can provide greater separation and resolution. After electrophoresis, the type and degree of protein phosphorylation can be investigated (Murthy and Iqbal, 1991; Gold *et al*., 1994), monosaccharide composition can be determined (Weitzhandler *et al*., 1993; Packer *et al*., 1995), and the structure and exact site of glycoamino acids can be investigated by either Edman degradation based techniques or by mass spectrometry (Pisano *et al*., 1993; Huberty *et al*., 1993; Carr, Huddleston and Bean, 1993). With further development of rapid techniques, investigation of phosphorylation and monosaccharides by chromatographic or mass spectrometric means is likely to become a routine step in the characterisation of post-translational modifications of proteins from reference maps.

The status of proteome projects

Many technical aspects of proteome research have already been discussed in this review, but an overview of the status of proteome projects has not yet been presented. Advances in proteome projects will initially rely on progress in genome sequencing initiatives, to enable an identity, amino acid sequence, or function to be assigned to each protein spot. *Table 7* shows genome size, proteome size, and the number of proteins already defined for a number of model organisms. This indicates that whilst genome sequencing programs for *E. coli* and *S. cerevisiae* are advanced, the massive size of some other genomes (and especially the human genome) means that their complete nucleotide sequences are unlikely to be available for many years. Because of this, 2-D reference maps and proteome projects of single cell organisms like *Mycoplasma sp.*, *E. coli* and *S. cerevisiae* will be the most detailed (Cordwell *et al.*, 1995; Wasinger *et al.*, 1995; Vanbogelen *et al.*, 1992; Garrels *et al.*, 1994), and complete maps of other organisms will take longer to construct. However, the use of cross-species protein identification techniques will allow proteomes of many prokaryotes and simple eukaryotes to be partially defined in reference to *E. coli* and *S. cerevisiae*.

Table 7: Estimated genome size, estimated proteome size, number of protein sequences in SWISS-PROT Release 31 (March, 1995), and approximate number of proteins of known identity on 2-D reference maps for some model organisms. Genome size data from Smith (1994), and total protein data from Bird (1995). Genome sequencing projects of *E. coli* and *S. cerevisiae* will probably be complete in 1996.

Species Name	Haploid genomesSize (million bp)	Estimated proteome size (total proteins)	Protein entries in SWISS PROT	Proteins annotated on 2-D Maps
Mycoplasma species	0.6–0.8	400–600	100	> 100
Escherichia coli	4.8	4000	3170	> 300
Saccharomyces cerevisiae	13.5	6000	3160	> 100
Dictyostelium discoideum	70	12500	204	-
Arabidopsis thaliana	70	14000	270	-
Caenorhabditis elegans	80	17800	703	-
Homo sapiens	2900	60000–80000	3326	> 1000

The study of vertebrate proteomes and vertebrate development is a phenomenal undertaking in comparison to the investigation of single cell organisms. This is because vast numbers of proteins are developmentally expressed, each body tissue has hundreds of unique proteins, and there are numerous tissue types. However, it is estimated that at least 35% of proteins in vertebrate cells will be conserved from tissue to tissue, constituting the 'housekeeping' proteins (Bird, 1995), with the remainder of proteins constituting a set that are specific to a cell type. Providing that standardised electrophoretic conditions are used, reference maps from many tissues of one organism can be superimposed in gel databases (e.g. Hochstrasser *et al.*, 1992). This accelerates the definition of the 'housekeeping' proteins, as well as sets of proteins that are unique to different tissue types. Such studies may, however, be complicated by post-translational modifications, which can differ on the same gene product in different tissues. Proteins that remain unknown after identification procedures will be useful in providing focus for nucleic acid sequencing initiatives.

FUTURE DIRECTIONS OF PROTEOME PROJECTS

This review has described recent advances in the area of proteome research. It has illustrated how new developments of older techniques (2-D electrophoresis and amino acid analysis) as well as the applications of new technology (mass spectrometry) have greatly widened the choice of tools the biologist and protein chemist has for the separation, identification and analysis of complex mixtures of proteins. This has made possible the establishment of detailed reference maps for organisms, which are becoming the method of choice for the definition of tissues or whole cells, and the investigation of gene expression therein.

Proteome projects are already impacting on the dogma of molecular biology that DNA sequence constitutes the definition of an organism. For example, the proteomes of different tissues of a single organism are often significantly different. Similarly, cross-species identification of proteins (for example the identification of proteins from *Candida albicans* by comparison with *S. cerevisiae*) can open up studies on organisms that are poorly molecularly defined. As cross-species identification can proceed at a pace orders of magnitude faster than a genome project in terms of defining the gene and protein complement of organims, the need for the DNA sequencing of genomes will be avoided, and emphasis placed on those found to be novel.

Just as genome sequencing is not an end in itself, neither is an annotated 2-D protein reference map of an organism, nor indeed the identification of proteins in a proteome. So whilst an immediate aim of proteome projects is to screen proteins in reference maps, this will lead to expression studies and characterisation of post-translational modifications. The challenge that then needs to be addressed is the investigation of structure and function of proteins in a proteome. The magnitude of this is illustrated by the fact that over half the open reading frames identified in *S. cerevisiae* chromosome III were initially of no known function (Oliver *et al.*, 1992). Structural and functional studies will be an undertaking just as formidable as genome studies are now and proteome projects are becoming, but will lead to an unimaginably detailed understanding of how living organisms are constructed and how they operate.

Acknowledgements

MRW is the recipient of an Australian Postgraduate Research Award. AAG, MRW, IHS and KLW acknowledge assistance for proteome projects through Macquarie University Research Grants, the Australian Research Council, the Australian National Health and Medical Research Council, Beckman Instruments and GBC Scientific Equipment. DH acknowledges the financial support of a Montus Foundation Grant and the Swiss National Fund for Scientific Research (Grant # 31-33658.92). We thank colleagues who supplied work that was 'In Press' during the writing of this review.

References

ANDERSON, N.L., HOFMANN, J.P., GEMMELL, A. AND TAYLOR, J. (1984). Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clinical Chemistry*, 30, 2031–2036.

ANDERSON, N.L. AND ANDERSON, N.G (1991). A two-dimensional gel database of human plasma proteins. *Electrophoresis*, 12, 883–906.
ANDERSON, N.L., ESQUER-BLASCO, R., HOFMANN, J.P. AND ANDERSON, N.G. (1991). A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis*, 12, 907–930.
ANDERSON, N.L., COPPLE, D.C., BENDELE, R.A., PROBST, G.S., RICHARDSON, F.C. (1992). Covalent protein modifications and gene expression changes in rodent liver following administration of methapyrilene: a study using two-dimensional electrophoresis. *Fundamental and Applied Toxicology*, 18, 570–580.
APPEL, R.D., BAIROCH, A. AND HOCHSTRASSER, D.F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends in Biochemical Sciences*, 19, 258–260.
APPEL, R.D., HOCHSTRASSER, D.F., FUNK, M., VARGAS, J.R., PELLEGRINI, C., MULLER, A.F. AND SCHERRER, J.R. (1991). The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis*, 12, 722–735.
APPEL, R.D., SANCHEZ, J-C., BAIROCH, A., GOLAZ, O., MIU, M., VARGAS, J.R. AND HOCHSTRASSER, D.F. (1993). SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis*, 14, 1323–1328.
APPEL, R.D., SANCHEZ, J-C., BAIROCH, A., GOLAZ, O., RAVIER, F., PASQUALI, C., HUGHES, G. AND HOCHSTRASSER, D.F. (1994). The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Research*, 22, 3581–3582.
BAIROCH, A. AND BOECKMANN, B. (1994). The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Research*, 22, 3578–3580.
BAKER, C.S., CORBETT, J.M., MAY, A.J., YACOUB, M.H. AND DUNN, M.J. (1992). A human myocardial two-dimensional electrophoresis database: protein characterisation by microsequencing and immunoblotting. *Electrophoresis*, 13, 723–726.
BARTLET-JONES, M., JEFFERY, W.A., HANSEN, H.F. AND PAPPIN, D.J.C. (1994). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Communications in Mass Spectrometry*, 8, 737–742.
BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOE, P. AND STRAUSS, M (1993). Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, 21, 4272–4280.
BERNERS-LEE, T.J., CAILLIAU, R., GROFF, J.F. AND POLLERMAN, B. (1992). *Electronic Networking: Research, Applications, and Policy*, 2, 52–58.
BIRD, A.P (1995) Gene number, noise reduction and biological complexity. *Trends in Genetics*, 11, 94–100
BJELLQVIST, B., EK, K., RIGHETTI, P.G., GIANAZZA, E., GORG, A., WESTERMEIER, R. AND POSTEL, W (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods*, 6, 317–339.
BJELLQVIST, B., PASQUALI, C., RAVIER, F., SANCHEZ, J-C AND HOCHSTRASSER, D.F. (1993a). A nonlinear wide-range immobilized pH gradient for two-dimensional electrophoresis and its definition in a relevant pH scale. *Electrophoresis*, 14, 1357–1365.
BJELLQVIST, B., SANCHEZ, J-C., PASQUALI, C., RAVIER, F., PAQUET, N., FRUTIGER, S., HUGHES, G.J AND HOCHSTRASSER, D.F. (1993b). Micropreparative 2-D electrophoresis allowing the separation of milligram amounts of proteins. *Electrophoresis*, 14, 1375–1378.
BJELLQVIST, B., HUGHES, G., PASQUALI, C., PAQUET, N., RAVIER, F., SANCHEZ, J-C., FRUTIGER, S. AND HOCHSTRASSER, D. (1993c). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14, 1023–1031.
BONNER, W.M. AND LASKEY, R.A. (1974). A film detection method for tritium-labeled proteins and nucleic acids in polyacrylamide gels. *European Journal of Biochemistry*, 46, 83–88.
BOUTELL, T., GARRELS, J.I., FRANZA, B.R., MONARDO, P.J. AND LATTER, G.I. (1994). REF52 on global gel navigator: an internet-accessible two-dimensional gel electrophoresis database. *Electrophoresis*, 15, 1487–1490.
BREWER, J., GRUND, E., HAGERLID, P., OLSSON, I. AND LIZANA, J. (1986). In *Electrophoresis '86* (M.J. Dunn, Ed.), pp. 226–229. VCH, Weinheim.

CARR S.A., HUDDLESTON, M.J. AND BEAN, M.F. (1993). Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Science*, 2, 183–196.

CAMPBELL, K.P., MACLENNAN, D.H. AND JORGENSEN, A.O. (1983). Staining of the Ca²⁺-binding proteins, calsequestrin, calmodulin, troponin C, and S-100, with the cationic dye "Stains-all". *Journal of Biological Chemistry*, 258, 11267–11273.

CELIS, J.E., CRUGER, D., KILL, J., DEJGAARD, K., LAURIDSEN, J.B., RATZ, G.P., BASSE, B., CELIS, A., RASMUSSEN, H.H., BAUW, G. AND VANDEKERKHOVE, J. (1990a). A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis*, 11, 242–254.

CELIS, J.E., GESSER, B., RASMUSSEN, H.H., MADSEN, P., LEFFERS, H., DEJGAARD, K., HONORE, B., OLSEN, E., RATZ, G., LAURIDSEN, J.B., BASSE, B., MOURITZEN, S., HELLERUP, M., ANDERSEN, A., WALBUM, E., CELIS, A., BAUW, G., PUYPE, M., VAN DAMME, J. AND VANDEKERKHOVE, J. (1990b). Comprehensive two-dimensional gel protein database offers a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis*, 11, 898–1071.

CELIS, J.E., RASMUSSEN, H.H., OLSEN, E., MADSEN, P., LEFFERS, H., HONORE, B., DEJGAARD, K., GROMOV, P., HOFFMANN, H.J., NIELSEN, M., VASSILEV, A., VINTERMYR, O., HAO, J., CELIS, A., BASSE, B., LAURIDSEN, J., RATZ, G.P., ANDERSEN, A.H., WALBUM, E., KJAERGAARD, I., PUYPE, M., VAN DAMME, J., AND VANDEKERKHOVE, J. (1993). The human keratinocyte two-dimensional database: update 1993. *Electrophoresis*, 14, 1091–1198.

CELIS, J.E. AND OLSEN, E. (1994). A qualitative and quantitative protein database approach identified individual and groups of functionally related proteins that are differentially regulated in simian virus 40 (SV40) transformed human keratinocytes: an overview of the functional changes associated with the transformed phenotype. *Electrophoresis*, 15, 309–344.

CORBETT, J.M., DUNN, M.J., POSCH, A. AND GORG, A. (1994a). Positional reproducibility of protein spots in two-dimsensional polyacrylamide electrophoresis using immobilzed pH gradient isolelectric focusing in the first dimension – an interlaboratory comparison. *Electrophoresis*, 15, 1205–1211.

CORBETT, J.M., WHEELER, C.H., BAKER, C.S., YACOUB, M.H. AND DUNN, M.J. (1994b). The human myocardial two-dimensional gel protein database: update 1994. *Electrophoresis*, 15, 1459–1465.

CORDWELL, S., WILKINS, M.R., CERPA-POLJAK, A., GOOLEY, A.A., DUNCAN, M., WILLIAMS, K.L AND HUMPHERY-SMITH, I. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using MALDI-TOF and amino acid composition. *Electrophoresis*, 15, 438–443.

COTTRELL, J.S. (1994). Protein identification by peptide mass fingerprinting. *Peptide Research*, 7, 115–124.

CRIMMINS, D.L., MCCOURT, D.W., THOMA, R.S., SCOTT, M.G., MACKE, K. AND SCHWARTZ, B.D. (1990). *In situ* chemical cleavage of proteins immobilized to glass-fiber and polyvinylidenefluoride membranes: cleavage at tryptophan residues with 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine to obtain internal amino acid sequence. *Analytical Biochemistry*, 187, 27–38.

DUTHEL, S. AND REVOL, A. (1993). Glycan microheterogeneity of alpha 1-antitrypsin in serum and meconium from normal and cystic fibrosis patients by crossed immunoaffinoelectrophoresis with different lectins (Con A, LCA, WGA). *Clinical and Chemical Acta*, 215, 173–187.

ECKERSKORN, C., JUNGBLUT, P., MEWES, W., KLOSE, J. AND LOTTSPEICH, F. (1988). Identification of mouse brain proteins after two-dimensional electrophoresis and electroblotting by microsequence analysis and amino acid composition. *Electrophoresis*, 9, 830–838.

ECKERSKORN, C., STRUPAT, K., KARAS, M., HILLENKAMP, F. AND LOTTSPEICH, F. (1992).

Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis*. 13. 661–665.

ECKERSKORN. C. AND LOTTSPEICH. F. (1993). Structural characterisation of blotting membranes and the influence of membrane parameters for electroblotting and subsequent amino acid sequence analysis of proteins. *Electrophoresis*. 14. 831–838.

EK. K.. BJELLQVIST. BJ AND RIGHETTI. P.G. (1983). Preparative isoelectric focusing in immobilized pH gradients. I General principles and methodology *Journal of Biochemical and Biophysical Methods*. 8. 135–155.

FEY. S.J.. CARLSEN. J.. MOSE LARSEN. P.. JENSEN. U.A.. KJELDSEN. K. AND HAUNSO. S. (1994). Two-dimensional gel electrophoresis as a tool for molecular cardiology. Proceedings of the International Society for Heart Research 'XV European Section Meeting'. pp 9–16

FREY. J.R.. KUHN. L.. KETTMAN. J.R AND LEFKOVITS. I. (1994). The amino acid composition of 350 lymphocyte proteins. *Molecular Immunology*. 31. 1219–1231.

GARRELS. J.I. (1989). The QUEST system for quantitative analysis of two-dimensional gels. *Journal of Biological Chemistry*. 264. 5269–5282.

GARRELS. J.I AND FRANZA. B.R. (1989) The REF52 protein database. *Journal of Biological Chemistry*. 264. 5283–5298.

GARRELS. J.I.. FRANZA. B.R.. CHANG. C. AND LATTER. G. (1990). Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation. serum stimulation. and viral transformation. *Electrophoresis*. 11. 1114–1130.

GARRELS. J.I.. FUTCHER. B.. KOBAYASHI. R.. LATTER. I.. SCHWENDER. B.. VOLPE. T.. WARNER. J.R. AND MCLAUGHLIN. C.S. (1994). Protein identification for a *Saccharomyces cerevisiae* protein database. *Electrophoresis*. 15. 1466–1486.

GELFI. C.. BOSSI. M.L.. BJELLQVIST. B. AND RIGHETTI. P.G. (1987). Isoelectric focusing in immobilized pH gradients in the pH 10–11 range. *Journal of Biochemical and Biophysical Research Methods*. 15. 41–48.

GHARAHDAGHI. F.. ATHERTON. D.. DEMOTT. M. AND MISCHE. S.M. (1992). Amino acid analysis of PVDF-bound proteins. in *Techniques in Protein Chemistry III* (R.H. Angeletti. Ed.). pp 249–260 Academic Press. San Diego.

GHOSH. P.. OKOH. C.. LIU. Q.H. AND LAKSHMAN. M.R. (1993). Effects of chronic ethanol on enzymes regulating sialylation and desialylation of transferrin in rats. *Alcoholism: Clinical and Experimental Research*. 17. 576–579.

GIOMETTI. C.S.. GEMMELL. M.A.. TOLLAKSEN. S.L. AND TAYLOR. J. (1991). Quantitation of human leukocyte proteins after silver staining: a study with two-dimensional electrophoresis. *Electrophoresis*. 12. 536–543

GIOMETTI. C.S.. TAYLOR. J. AND TOLLAKSEN. S.L. (1992). Mouse liver protein database: a catalog of proteins detected by two-dimensional gel electrophoresis. *Electrophoresis*. 13. 970–991.

GOLAZ. O. HUGHES. G.J.. FRUTIGER. S.. PAQUET. N.. BAIROCH. A.. PASQUALI. C.. SANCHEZ. J.C.. TISSOT. J.D.. APPEL. R.D.. WALZER. C.. BALANT. L. AND HOCHSTRASSER. D.F (1993) Plasma and red blood cell protein maps: update 1993 *Electrophoresis*. 14. 1223–1231.

GOLD. M.R.. YUNGWIRTH. T.. SUTHERLAND. C.L.. INGHAM. R.J.. VIANZON. D.. CHIU. R.. VAN-OOSTVEEN. I.. MORRISON. H.D. AND AEBERSOLD. R. (1994). Purification and identification of tyrosine-phosphorylated proteins from lymphocytes stimulated through the antigen receptor. *Electrophoresis*. 15. 441–453.

GOLDBERG. H.A.. DOMENICUCCI. C.. PRINGLE. G.A. AND SODEK. J. (1988). Mineral-binding proteoglycans of fetal porcine calvarial bone. *Journal of Biological Chemistry*. 263. 12092–12101.

GOOLEY. A.A.. MARSHCHALEK. R AND WILLIAMS. K.L. (1992). Size polymorphisms due to changes in the number of *O*-glycosylated tandem repeats in the *Dictyostelium discoideum* glycoprotein PsA. *Genetics*. 130. 749–756.

GORG. A.. POSTEL. W. AND GUNTHER. S. (1988). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*. 9. 531–546.

GORG. A.. POSTEL. W.. GUNTHER. S.. WESER. J.. STRAHLER. J.R.. HANASH. S.M.. SOMERLOT. L.

AND KLICK, R. (1988). Approach to stationary two-dimensional pattern: influence of focusing time and immobiline/carrier ampholyte concentrations. *Electrophoresis*, 9, 37–46.

GRAVEL, P., GOLAZ, O., WALZER, C., HOCHSTRASSER, D.F., TURLER, H., AND BALANT, L.P. (1994). Analysis of glycoproteins separated by two-dimensional gel electrophoresis using lectin blotting revealed by chemiluminescence. *Analytical Biochemistry*, 221, 66–71.

GUNTHER, S., POSTEL, W., WIERING, H. AND GORG, A. (1988). Acid phosphatase typing for breeding nematode-resistant tomatoes by isoelectric focusing with an ultranarrow immobilized pH gradient. *Electrophoresis*, 9, 618–620.

HANASH, S.M., STRAHLER, J.R., NEEL, J.V., HAILAT, N., MELHEM, R., KEIM, D., ZHU, X.X., WAGNER, D., GAGE, D.A. AND WATSON, J.T. (1991). Highly resolving two-dimensional gels for protein sequencing. *Proceedings of the National Academy of Sciences USA*, 88, 5709–5713.

HARRINGTON, M.G., COFFMAN, J.A., CALZONE, F.J., HOOD, L.E., BRITTEN, R.J. AND DAVIDSON, E.H. (1992). Complexity of sea urchin embryo nuclear proteins that contain basic domains. *Proceedings of the National Academy of Sciences USA*, 89, 6252–6256.

HARRINGTON, M.G., LEE, K.H., YUN, M., ZEWERT, T., BAILEY, J.E. AND HOOD, L.E. (1993). Mechanical precision in two-dimensional electrophoresis can improve spot positional reproducibility. *Applied and Theoretical Electrophoresis*, 3, 347–353.

HELLMAN, U., WERNSTEDT, C., GONEZ, J. AND HELDIN, C-H. (1995). Improvement of an in-gel digestion for the micropreparation of internal protein fragments for amino acid sequencing. *Analytical Biochemistry*, 224, 451–455.

HENZEL, W.J., BILLECI, T.M., STULTS, J.T., WONG, S.C., GRIMLEY, C. AND WATANABE, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences USA*, 90, 5011–5015.

HESS, D., COVEY, T.C., WINZ, R., BROWNSEY, R.W. AND AEBERSOLD, R. (1993). Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Science*, 2, 1342–1351.

HOBOHM, U., HOLTHAEVE, T. AND SANDER, C. (1994). Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Analytical Biochemistry*, 222, 202–209.

HOCHSTRASSER, D.F. AND MERRIL, C.R. (1988). 'Catalysts' for polyacrylamide gel polymerization and detection of proteins by silver staining. *Applied and Theoretical Electrophoresis*, 1, 35–40.

HOCHSTRASSER, D.F., PATCHORNIK, A., AND MERRIL, C.R. (1988). Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. *Analytical Biochemistry*, 173, 412–423.

HOCHSTRASSER, A.C., JAMES, R.W., POMETTA, D. AND HOCHSTRASSER, D.F. (1991a). Preparative isoelectrofocusing and high resolution two-dimensional electrophoresis for concentration and purification of proteins. *Applied and Theoretical Electrophoresis*, 1, 333–337.

HOCHSTRASSER, D.F., APPEL, R.D., VARGAS, R., PERRIER, R., VURLOD, J.F., RAVIER, F., PASQUALI, C., FUNK, M., PELLIGRINI, C., MULLER, A.F. AND SCHERRER, J.R. (1991b). A clinical molecular scanner: the Melanie project. *Medical Computing*, 8, 85–91.

HOCHSTRASSER D.F., FRUTIGER, S., PAQUET, N., BAIROCH, A., RAVIER, F., PASQUALI, C., SANCHEZ, J-C., TISSOT, J-D., BJELLQVIST, B., VARGAS, R., APPEL, R.D. AND HUGHES, G.J. (1992). Human liver protein map: a reference database established by microsequencing and gel comparison. *Electrophoresis*, 13, 992–1001.

HOLT, T.G., CHANG, C., LAURENT-WINTER, C., MURAKAMI, T., DAVIES, J.E. AND THOMPSON, C.J. (1992). Global changes in gene expression related to antibiotic synthesis in *Streptomyces hygroscopicus*. *Molecular Microbiology*, 6, 969–980.

HONORE, B., LEFFERS, H., MADSEN, P. AND CELIS, J.E. (1993). Interferon-gamma up-regulates a unique set of proteins in human keratinocytes. Molecular cloning and expression of the cDNA encoding the RGD-sequence containing protein IGUP I-5111. *European Journal of Biochemistry*, 218, 421–430.

HUBERTY, M.C., VATH, J.E., YU, W. AND MARTIN, S.A. (1993). Site-specific carbohydrate

identification in recombinant proteins using MALD-TOF MS. *Analytical Chemistry*, 65, 2791–2800.

HUGHES, G.J., FRUTIGER, S., PAQUET, N., PASQUALI, C., SANCHEZ, J-C., TISSOT, J.D., BAIROCH, A., APPEL, R.D. AND HOCHSTRASSER, D.F. (1993). Human liver protein map update 1993. *Electrophoresis*, 14, 1216–1222.

HUGHES, J.H., MACK, K. AND HAMPARIAN, V.V. (1988). India ink staining of proteins on nylon and hydrophobic membranes. *Analytical Biochemistry*, 173, 18–25.

JAMES, P., QUADRONI, M., CARAFOLI, E. AND GONNET, G. (1993). Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, 195, 58–64.

JI, H., WHITEHEAD, R.H., REID, G.E., MORITZ, R.L., WARD, L.D. AND SIMPSON, R.J. (1994). Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. *Electrophoresis*, 15, 391–405.

JOHNSON, R.S. AND WALSH, K.A. (1992). Sequence analysis of peptide mixtures by automated integration of Edman and mass spectrometric data. *Protein Science*, 1, 1083–1091.

JOHNSTON, R.F., PICKETT, S.C. AND BARKER, D.L. (1990). Autoradiography using storage phosphor technology. *Electrophoresis*, 11, 355–360.

JUNGBLUT, P., DZIONARA, M., KLOSE, J. AND WITTMANN-LEIBOLD, B. (1992). Identification of tissue proteins by amino acid analysis after purification by two-dimensional electrophoresis. *Journal of Protein Chemistry*, 11, 603–612.

JUNGBLUT, P., OTTO, A., ZEINDL-EBERHART, E., PLEIßNER, K-P., KNECHT, M., REGITZ-ZAGROSEK, V., FLECK, E. AND WITTMANN-LEIBOLD, B. (1994). Protein composition of the human heart: the construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis*, 15, 685–707.

KOHARA, Y., AKIYAMA, K. AND ISONO, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, 50, 495–508.

KLOSE, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for individual point mutations in mammals. *Human Genetik*, 26, 231–243.

LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1991). Quantitative analysis of protein synthesis in mouse embryos I: extensive re-programming at the one- and two-cell stages. *Development*, 2, 921–932.

LATHAM, K.E., GARRELS, J.I., CHANG, C. AND SOLTER, D. (1992). Analysis of embryonic mouse development: construction of a high-resolution, two-dimensional gel protein database. *Applied and Theoretical Electrophoresis*, 2, 163–170.

LATHAM, K.E., GARRELS, J.I. AND SOLTER, D. (1993). Two-dimensional analysis of protein synthesis. *Methods in Enzymology*, 255, 473–489.

LE MAIRE, M., DESCHAMPS, S., MOLLER, J.V., LE CAER, J.P. AND ROSSIER, J. (1993). Electrospray ionization mass spectrometry from sodium dodecyl sulfate-polyacrylamide gel electrophoresis: application to the topology of the sarcoplasmic reticulum Ca²⁺-ATPase. *Analytical Biochemistry*, 214, 50–57.

LEMKIN, P.F. AND LESTER, E.P. (1989). Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modelling. *Electrophoresis*, 10, 122–140.

LEMKIN, P.F., WU, Y. AND UPTON, K. (1993). An efficient disk-based data structure for rapid searching of quantitative two-dimensional gel databases. *Electrophoresis*, 14, 1341–1350.

LI, K.W., GERAERTS, W.P., VAN-ELK, R. AND KOOSE, J. (1989). Quantification of proteins in the subnanogram and nanogram range: comparison of the AuroDye, FerriDye, and india ink staining methods. *Analytical Biochemistry*, 182, 44–47.

LIANG, P. AND PARDEE, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257, 967–971.

MANN, M. (1995). Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (R. Kellner, F. Lottspeich, and H.E. Meyer, Eds), pp 223–245. VCH, Weinheim.

MANN, M., HOJRUP, P. AND ROEPSTORFF, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, 22, 338–345.

MANN, M. AND WILM, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66, 4390–4399.

MATSUDAIRA, P. (1987). Sequence of picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *Journal of Biological Chemistry*, 262, 10035–10038.

MONARDO, P.J., BOUTELL, T., GARRELS, J.I. AND LATTER, G.I. (1994). A distributed system for two-dimensional gel analysis. *Computer Applications in the Biosciences*, 10, 137–143.

MORTZ, E., VORM, O., MANN, M. AND ROEPSTORFF, P. (1994). Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biological Mass Spectrometry*, 23, 249–261.

MURTHY, L.R. AND IQBAL, K. (1991). Measurement of picomoles of phosphoamino acids by high performance liquid chromatography. *Analytical Biochemistry*, 193, 299–305.

MYRICK, J.E., LEMKIN, P.F., ROBINSON, M.K. AND UPTON, K.M. (1993). Comparison of the BioImage Visage 2000 and the GELLAB-II two-dimensional electrophoresis image analysis systems. *Applied and Theoretical Electrophoresis*, 3, 335–346.

NEIDHARDT, F.C., APPLEBY, D.B., SANKAR, P., HUTTON, M.E. AND PHILLIPS, T.A. (1989). Genomically linked cellular protein databases derived from two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis*, 10, 116–122.

NIKODEM, V. AND FRESCO, J.R. (1979). Protein fingerprinting by SDS-gel electrophoresis after partial fragmentation with CNBr. *Analytical Biochemistry*, 97, 382–386.

NOKIHARA, K., MORITA, N. AND KURIKI, T. (1992). Applications of an automated apparatus for two-dimensional electrophoresis, Model TEP-1, for microsequence analysis of proteins. *Electrophoresis*, 13, 701–707.

O'FARRELL, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250, 4007–4021.

O'FARRELL, P.Z., GOODMAN, H.M. AND O'FARRELL, P.H. (1977). High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell*, 12, 1133–1142.

OLIVER *et al.* (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46.

OLSEN, A.D. AND MILLER, M.J. (1988). Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Analytical Biochemistry*, 169, 49–70.

ORTIZ, M.L., CALERO, M., FERNANDEZ-PATRON, C., PATRON, C.F., CASTELLANOS, L. AND MENDEZ, E. (1992). Imidazole-SDS-Zn reverse staining of proteins in gels containing or not SDS and microsequence of individual unmodified electroblotted proteins. *FEBS Lettersers*, 296, 300–304.

OSTERGREN, K., ERIKSSON, G. AND BJELLQVIST, B. (1988). The influence of support material used on band sharpness in Immobiline gels. *Journal of Biochemical and Biophysical Methods*, 16, 165–170.

OU, K., WILKINS, M.R., YAN, J.X., GOOLEY, A.A., FUNG, Y., SHEUMACK, D. AND WILLIAMS, K.L. (1995). Improved high-performance liquid chromatography of amino acids derivatised with 9-fluorenylmethyl chloroformate. *Journal of Chromatography* (in press).

PACKER, N., WILKINS, M.R., GOLAZ, O., LAWSON, M., GOOLEY, A.A., HOCHSTRASSER, D.F., REDMOND, J. AND WILLIAMS, K.L. (1995). Characterisation of human plasma glycoproteins separated by two-dimensional gel electrophoresis. *Bio/Technology* (in press).

PAPPIN, D.J.C., HOJRUP, P. AND BLEASBY, A.J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3, 327–332.

PATTERSON, S.D. (1994). From electrophoretically separated protein to identification: strategies for sequence and mass analysis. *Analytical Biochemistry*, 221, 1–15.

PATTERSON, S.D. AND LATTER, G.I. (1993). Evaluation of storage phosphor imaging for quantitative analysis of 2-D gels using the Quest II system. *BioTechniques*, 15, 1076–1083.

PISANO, A., REDMOND, J.W., WILLIAMS, K.L. AND GOOLEY, A.A. (1993). Glycosylation sites identified by solid-phase Edman degradation: *O*-linked glycosylation motifs on human glycophorin A. *Glycobiology*, 3, 429–435.

RABILLOUD, T. (1992). A comparison between low background silver diamine and silver nitrate protein stains. *Electrophoresis*, 13, 429–439.

RASMUSSEN, H.H., VAN DAMME, J., PUYPE, M., GESSER, B., CELIS, J.E. AND VANDEKERCKHOVE, J. (1992). Microsequences of 145 proteins recorded in the two-dimensional gel protein database of normal human epidermal keratinocytes. *Electrophoresis*, 13, 960–969.

RASMUSSEN, H.H., MORTZ, E., MANN, M., ROEPSTORFF, P. AND CELIS, J.E. (1994). Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass-spectrometric peptide mapping alone and in combination with microsequencing. *Electrophoresis*, 15, 406–416.

RICHARDSON, F.C., HORN, D.M. AND ANDERSON, N.L. (1994). Dose-responses in rat hepatic protein modification and expression following exposure to the rat hepatocarcinogen methapyrilene. *Carcinogenesis*, 15, 325–329.

RIGHETTI, P.G. (1990). Immobilized pH gradients: theory and methodology. In *Laboratory Techniques in Biochemistry and Molecular Biology* (R.H. Burdon and P.H. van Knippenberg, Eds) Elsevier, Amsterdam.

RIGHETTI, P.G. AND DRYSDALE, J.W. (1973). *Annals of the New York Academy of Sciences*, 209, 163–186.

RODRIGUEZ, L.V., GERNSTEN, D.M., RAMAGLI, L.S. AND JOHNSTON, D.A. (1993). Towards stoichiometric silver staining of proteins resolved in complex two-dimensional electrophoresis gels: real-time analysis of pattern development. *Electrophoresis*, 14, 628–637.

ROSENFELD, J., CAPDEVIELLE, J., GUILLEMOT, J.C. AND FERRARA, P. (1992). In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical Biochemistry*, 203, 173–179.

SANCHEZ, J.C., RAVIER, F., PASQUALI, C., FRUTIGER, S., PAQUET, N., BJELLQVIST, B., HOCHSTRASSER, D.F. AND HUGHES, G.J. (1992). Improving the detection of proteins after transfer to polyvinylidene difluoride membranes. *Electrophoresis*, 13, 715–717.

SANGER, F., COULSON, A.R., HONG, G.F., HILL, D.F. AND PETERSEN, G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162, 729–773.

SCHEELE, G.J. (1975). Two-dimensional analysis of soluble proteins. *Biochemistry*, 250, 5375–5385.

SHAW, G. (1993). Rapid identification of proteins. *Proceedings of the National Academy of Sciences USA*, 90, 5138–5142.

SIBBALD, P.R., SOMMERFELDT, H. AND ARGOS, P. (1991). Identification of proteins in sequence databases from amino acid composition. *Analytical Biochemistry*, 198, 330–333.

SIMPSON, R.J., TSUGITA, A., CELIS, J.E., GARRELS, J.I. AND MEWES, H.W. (1992). Workshop on two-dimensional gel protein databases. *Electrophoresis*, 13, 1055–1061.

SINHA, P.K., KOTTGEN, E., STOFFLER, M-M., GIANAZZA, E. AND RIGHETTI, P.G. (1990). Two-dimensional maps in very acidic immobilized pH gradients. *Journal of Biochemical and Biophysical Methods*, 20, 345–352.

SMITH, D.W. (1994). Introduction. In *Biocomputing: Informatics and Genome Projects* (D.W. Smith, Ed.), pp1–12. Academic Press, San Diego.

STRUPAT, K., KARAS, M., HILLENKAMP, F., ECKERSKORN, C. AND LOTTSPEICH, F. (1994). Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Analytical Chemistry*, 66, 464–470.

SUTTON, C.W., PEMBERTON, K.S., COTTRELL, J.S., CORBETT, J.M., WHEELER, C.H., DUNN, M.J. AND PAPPIN, D.J. (1995). Identification of myocardial proteins from two-dimensional gels by peptide mass fingerprinting. *Electrophoresis*, 16, 308–316.

TOUS, G.I., FAUSNAUGH, J.L., AKINYOSOYE, O., LACKLAND, H., WINTERCASH, P., VITORICA, F.J. AND STEIN, S. (1989). Amino acid analysis on polyvinylidene difluoride membranes. *Analytical Biochemistry*, 179, 50–55.

TOVEY, E.R., FORD, S.A. AND BALDO, B.A. (1987). Protein blotting on nitrocellulose: some important aspects of the resolution and detection of antigens in complex extracts. *Journal of Biochemical and Biophysical Methods*, 14, 1–17.

URWIN, V.E. AND JACKSON, P. (1993). Two-dimensional polyacrylamide gel electrophoresis of proteins labeled with the fluorophore monobromobimane prior to first-dimensional isoelectric focusing: imaging of the fluorescent protein spot patterns using a cooled charge-coupled device. *Analytical Biochemistry*, 209, 57–62.

VANBOGELEN, R.A., HUTTON, M.E. AND NEIDHARDT, F.C. (1990). Gene-protein database

of *Escherichia coli* K-12: edition 3. *Electrophoresis*. **11**. 1131–1166.

VANBOGELEN, R.A. AND NEIDHARDT, F.C. (1991). The gene-protein database of *Escherichia coli*: edition 4. *Electrophoresis*. **12**. 955–994.

VANBOGELEN, R.A., SANKER, P., CLARK, R.L., BOGAN, J.A. AND NEIDHARDT, F.C. (1992). The gene-protein database of *Escherichia coli*: edition 5. *Electrophoresis*. **13**. 1014–1054.

VANDEKERKHOVE, J., BAUW, G., VANCOMPERNOLLE, K., HONORE, B. AND CELIS, J. (1990). Comparative two-dimensional gel analysis and microsequencing identifies gelsolin as one of the most prominent downregulated markers of transformed human fibroblast and epithelial cells. *Journal of Cell Biology*. **111**. 95–102.

VANFLETEREN, J.R., RAYMACKERS, J.G., VAN BUN, S.M. AND MEHUS, L.A. (1992). Peptide mapping and microsequencing of proteins separated by SDS-PAGE after limited *in situ* hydrolysis. *BioTechniques*. **12**. 550–557.

VORM, O. AND MANN, M. (1994). Improved mass accuracy in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of peptides. *Journal of the American Society for Mass Spectrometry*. **5**. 955–958.

VORM, O., ROEPSTORFF, P. AND MANN, M. (1994). Improved resolution and very high sensitivity in MALDI TOF of matrix surfaces made by fast evaporation. *Analytical Chemistry*. **66**. 3281–3287.

WALLACE, A. AND SALUZ, H.P. (1992a). Ultramicrodetection of proteins in polyacrylamide gels. *Analytical Biochemistry*. **203**. 27–34.

WALLACE, A. AND SALUZ, H.P. (1992b). Beyond silver staining. *Nature*. **357**. 608–609.

WALSH, B.J., GOOLEY, A.A., WILLIAMS, K.L. AND BREIT, S.N. (1995). Identification of macrophage activation associated proteins by two-dimensional electrophoresis and microsequencing. *Journal of Leukocyte Biology*. **57**. 507–512.

WASINGER, V.C., CORDWELL, S.J., POLJAK, A., YAN, J.X., GOOLEY, A.A., WILKINS, M.R., DUNCAN, M., HARRIS, R., WILLIAMS, K.L. AND HUMPHERY-SMITH, I. (1995). Progress with Gene-Product Mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*. **16**. In Press.

WEITZHANDLER, M., KADLECEK, D., AVDALOVIC, N., FORTE, J.G., CHOW, D. AND TOWNSEND, R.R. (1993). Monosaccharide and oligosaccharide analysis of proteins transferred to polyvinylidene fluoride membranes after sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry*. **268**. 5121–5130.

WILKINS, M.R., PASQUALI, C., APPEL, R.D., OU, K., GOLAZ, O., SANCHEZ, J.-C., YAN, J.X., GOOLEY, A.A., HUGHES, G., HUMPHERY-SMITH, I., WILLIAMS, K.L., AND HOCHSTRASSER, D.F. (1995). From Proteins to Proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Submitted.

WILKINS, M.R., OU, K., APPEL, R.D., GOLAZ, O., PASQUALI, C., YAN, J.X., FARNSWORTH, V., CARTIER, P., HOCHSTRASSER, D.F., WILLIAMS, K.L. AND GOOLEY, A.A. (1996). Rapid protein identification using N-terminal sequence tagging and amino acid analysis (submitted).

WIRTH, P.J., LUO, L.D., FUJIMOTO, Y., BISGAARD, H.C. AND OLSEN, A.D. (1991). The rat liver epithelial (RLE) cell protein database. *Electrophoresis*. **12**. 931–954.

WIRTH, P.J., LUO, L.D., BENJAMIN, T., HOANG, T.N., OLSEN, A.D. AND PARMALEE, D.C. (1993). The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis*. **14**. 1199–1215.

WU, Y., LEMKIN, P.F. AND UPTON, K. (1993). A fast spot segmentation algorithm for two-dimensional gel electrophoresis analysis. *Electrophoresis*. **14**. 1351–1356.

YAMAGUCHI, K. AND ASAKAWA, H. (1988). Preparation of colloidal gold for staining proteins electrotransferred onto nitrocellulose membranes. *Analytical Biochemistry*. **172**. 104–107.

YAMASHITA, K., IDEO, H., OHKURA, T., FUKUSHIMA, K., YUASA, I., OHNO, K. AND TAKESHITA, K. (1993). Sugar chains of serum transferrin from patients with carbohydrate deficient glycoprotein syndrome. Evidence of asparagine-N-linked oligosaccharide transfer deficiency. *Journal of Biological Chemistry*. **268**. 5783–5789.

YATES, J.R. III, SPEICHER, S., GRIFFIN, P.R. AND HUNKAPILLER, T. (1993). Peptide mass maps: a highly informative approach to protein identification. *Analytical Biochemistry*. **214**. 397–408.

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref. No. D

Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing

JULIO E. CELIS,*,[1] HANNE H. RASMUSSEN,* HENRIK LEFFERS,* PEDER MADSEN,* BENT HONORÉ,* BORBALA GESSER,* KURT DEJGAARD,* JOËL VANDEKERCKHOVE†

*Institute of Medical Biochemistry and Human Genome Research Centre, Aarhus University, DK-8000 Aarhus, Denmark and †Laboratorium voor Fysiologische Chemie, Rijksuniversiteit Gent, Belgium

ABSTRACT **Analysis of cellular protein patterns by computer-aided 2-dimensional gel electrophoresis together with recent advances in protein sequence analysis have made possible the establishment of comprehensive 2-dimensional gel protein databases that may link protein and DNA information and that offer a global approach to the study of the cell. Using the integrated approach offered by 2-dimensional gel protein databases it is now possible to reveal phenotype specific protein (or proteins), to microsequence them, to search for homology with previously identified proteins, to clone the cDNAs, to assign partial protein sequence to genes for which the full DNA sequence and the chromosome location is known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Human 2-dimensional gel protein databases are becoming increasingly important in view of the concerted effort to map and sequence the entire genome.——Celis, J. E.; Rasmussen, H. H.; Leffers, H.; Madsen, P.; Honoré, B.; Gesser, B.; Dejgaard, K.; Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB J.* 5: 2200–2208; 1991.**

Key Words: human protein patterns · 2-dimensional gel protein databases · gene expression · microsequencing · cDNA cloning · linking protein and DNA information · genome mapping and sequencing

PROTEINS SYNTHESIZED FROM information contained in the DNA orchestrate most cellular functions. The total number of proteins synthesized by a typical human cell is unknown although current estimates range from 3000 to 6000. Of these, as many as 70% may perform household functions and are expected to be shared by all cell types irrespective of their origin. There are many different cell types in the human body with perhaps 30,000 to 50,000 proteins expressed in the organism as a whole judged from the fact that about 3% of the haploid genome correspond to genes. Today only a small fraction of the total set of proteins has been identified, and little is known about the protein patterns of individual cell types or their variation under physiological and abnormal conditions.

For the past 15 years, high resolution 2-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (refs 1–6 and references therein). The technique originally described by O'Farrell separates proteins in terms of their isoelectric point (pI) and molecular weight. Usually one chooses a condition of interest and the cell reveals the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available 2-dimensional gel techniques (regular gel format) can resolve between 1000 and 2000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by analyzing partiall purified cellular fractions.

Two-dimensional gel ectrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (refs 1–6, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems and the lack of means for storing and communicating protein information. Only recently, because of the development of appropriate computer software (7–13), has it been possible to scan gels, assign numbers to individual proteins, and store the wealth of information in quantitative and qualitative comprehensive 2-dimensional gel protein databases (4, 14–23), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type. Such integrated 2-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to study the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation, and diagnosis of genetic and clinical diseases (**Fig. 1**).

The approach using systematic 2-dimensional gel protein analysis has recently gained a new dimension with the advent of techniques to microsequence major proteins recorded

[1]To whom correspondence should be addressed, at: Institute of Medical Biochemistry and Human Genome Research Centre, Ole Worms Alle, Bldg. 170, University Park, DK-8000 Aarhus C, Denmark.

0892-6638/91/0005-2200/$01.50. © FASEB

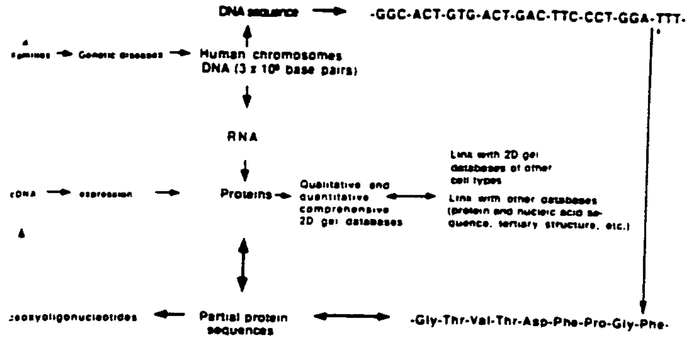


Figure 1. Interface between partial protein sequence databases. comprehensive 2-dimensional gel databases. and the human genome sequencing project. Appropriate software is required to compare protein and DNA sequences. In general. although the inference of a protein's sequence from the DNA sequence (thick arrow) is direct and unambiguous. the DNA sequence can only be inferred approximately from the protein sequence (thin arrow) and cloning of the gene requires either a cDNA or the requisite group of oligonucleotide probes deduced from the partial amino acid sequence. Modified from ref 6.

in the databases (refs 24–42 and references therein). Partial protein sequences can be used to search for protein identity as well as to prepare specific DNA probes for cloning as-yet-uncharacterized proteins (Fig. 1). As these sequences can be stored in the database (see for example Fig. 2H), they offer a unique opportunity to link information on proteins with the existing or forthcoming DNA sequence data on the human genome (Fig. 1) (20, 36, 39).

Using the integrated approach offered by comprehensive 2-dimensional gel databases (Fig. 1), it will be possible to identify phenotype-specific proteins: microsequence them and store the information in the database: search for homology with previously characterized proteins; clone the cDNAs, assign partial protein sequences to genes for which the full DNA sequence and the chromosome location are known, and study the regulatory properties and function of groups of proteins (pathways, organelles, etc.) that are coordinately expressed in a given biological process. Comprehensive 2-dimensional gel protein databases will depict an integrated picture of the expression levels and properties of the thousands of protein components of organelles, pathways, and cytoskeletal systems in both physiological and abnormal conditions and are expected to lead to identification of new regulatory networks in different cell types and organisms. In the future, 2-dimensional gel protein databases may be linked to each other as well as to national and international specialized databanks on nucleic acid and protein sequences, protein structures, NMR experimental data, complex carbohydrates, etc.

A few 2-dimensional gel protein databases that are accessible in a computer form have been published in extenso: these correspond to the protein-gene database of *Escherichia coli* K-12 developed by Neidhardt and colleagues (14, 23), the rat REF 52 database established by Garrels and co-workers at Cold Spring Harbor (18, 22), and a few human databases (transformed amnion cells [15, 20], normal embryonal lung MRC-5 fibroblasts [17, 21]; keratinocytes [19] and peripheral blood mononuclear cells [15]) developed in Aarhus. Given space limitations and to keep this review in focus, we will concentrate on the computerized analysis of human cellular 2-dimensional gel patterns, and in particular on the steps involved in establishing comprehensive 2-dimensional gel databases that can link protein and DNA information.

MAKING AND MANAGING A COMPREHENSIVE 2-DIMENSIONAL GEL DATABASE OF HUMAN CELLULAR PROTEINS

The first step in making a comprehensive 2-dimensional gel protein database is to prepare a synthetic image (digital form of the gel image) of the gel (fluorogram, Coomassie blue or silver stained gel) to be used as a standard or master reference. This can be done with laser scanners, charge couple device (CCD)[2] array scanners, television cameras, rotating drum scanners, and multiwire chambers (13). Computerized analysis systems for spot detection, quantitation, pattern matching, and data handling (access and retrieval of information, database making) have been described in the literature (ELSIE [43], GELLAB [11], HERMeS [44], MELANIE [10], QUEST (9), and TYCHO [8]) and some are available commercially (PDQUEST, Protein Database Inc., Huntington, N.Y.; KEPLER, Large Scale Biology, Rockville, Md.; Visage, BioImage Corporation, Ann Arbor, Mich.; Gemini, Joyce Loebl, Gateshead; Microscan 1000, Technology Resources Inc., Nashville, Tenn. and MasterScan, Billerica, Mass.). Unfortunately, most of these systems are incompatible with one another and their advantages and disadvantages have been discussed by Miller (13).

In our work station in Aarhus, fluorograms are scanned with a Molecular Dynamics laser scanner and the data are analyzed using the PDQUEST II software (Protein Databases Inc.) (12) running on a spark station computer 4100 FC-8-P3 from SUN Microsystems, Inc. The scanner measures intensity in the range of 0–2.0 absorbance. A typical scan of a 17 × 17 cm fluorogram takes about 2 min. Steps in image analysis include: initial smoothing, background substraction, final smoothing, spot detection, and fitting of ideal Gaussian distribution to spot centers. Spot intensity is calculated as the integration of a fitted Gaussian. If calibration strips containing individual segments of a known amount of radioactivity are used, it is possible to merge multiple exposures of the sample image into a single data image of greater dynamic range. Once the synthetic image is created it can be stored on disk and displayed directly on the monitor. Functions that can be used to edit the images include: cancel (for example, to erase scratches that may have been interpreted as spots by the computer; cancel streaks or low dpm spots), combine (sometimes a spot may be resolved into several closely packed spots), restore, uncombine, and add spot to the gel. The process is time consuming—about 1-1/2 day per image. Edited standard images can be matched to other synthetic images. **Figure 2A** shows a portion of a standard synthetic image (IEF) of a fluorogram of [^{35}S]methionine labeled cellular proteins from human AMA cells (master database) (20). Images can be displayed either in black and white (resembling the original fluorograms) or in color (other images in Fig. 2), depending on the need. As shown in Fig. 2*B*, each polypeptide is assigned a number by the computer, which facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel (20). The standard image can be matched automatically by the computer to other standard or reference gels (Fig. 2*C*, matching of AMA cellular proteins [left] to MRC-5 proteins [right]) provided a few landmark spots are given manually as reference (indicated with a + in Fig. 2*C*) to initiate the process.

---

[2]Abbreviations: CCD, charge couple device; PCNA, proliferating cell nuclear antigen; HPLC, high performance liquid chromatography.

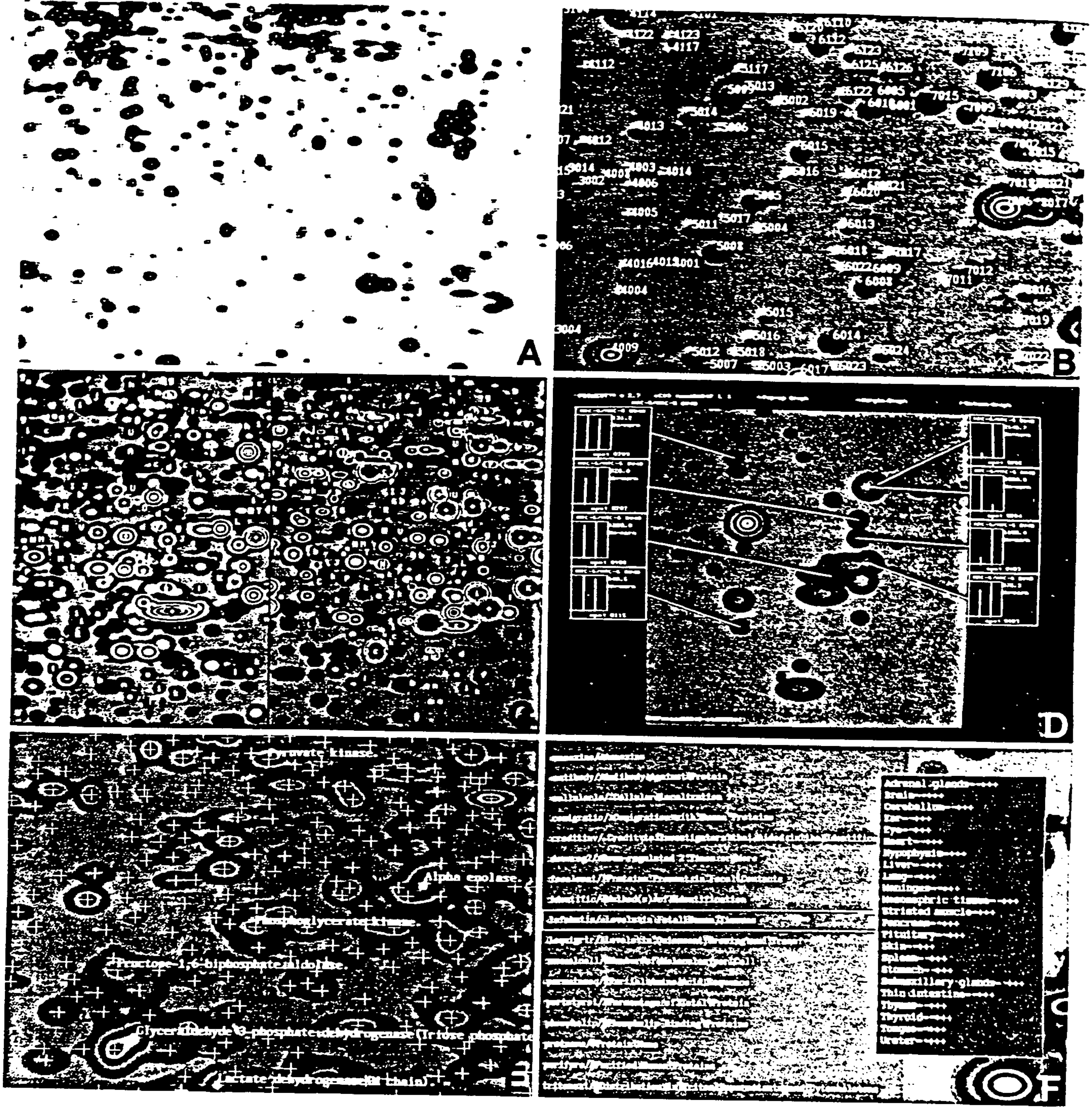
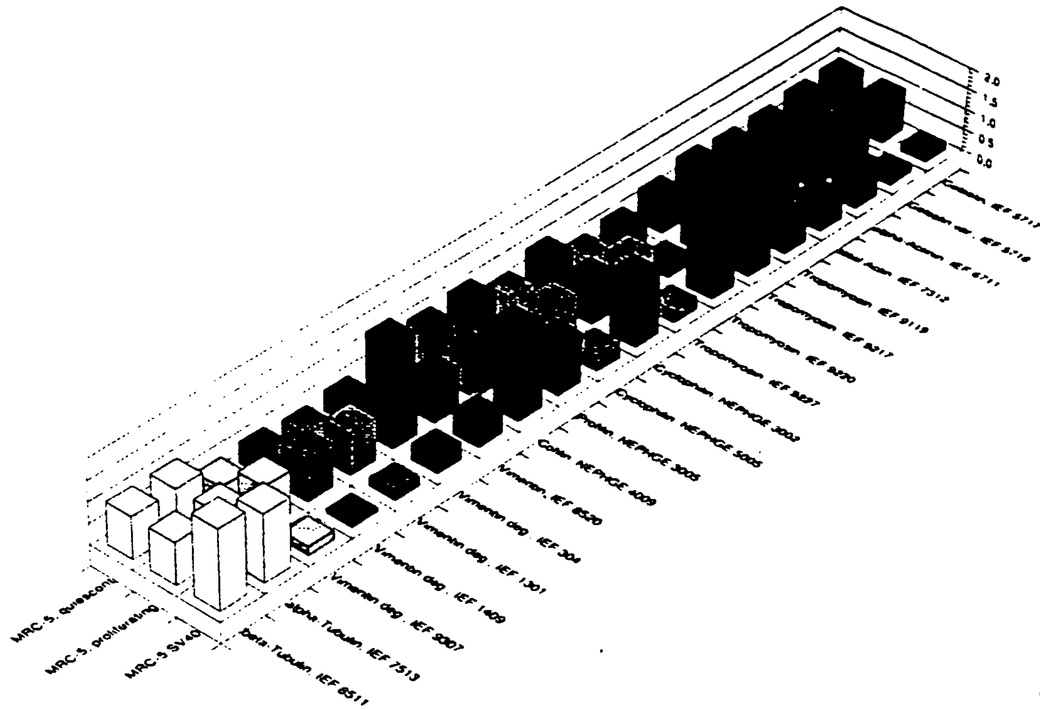
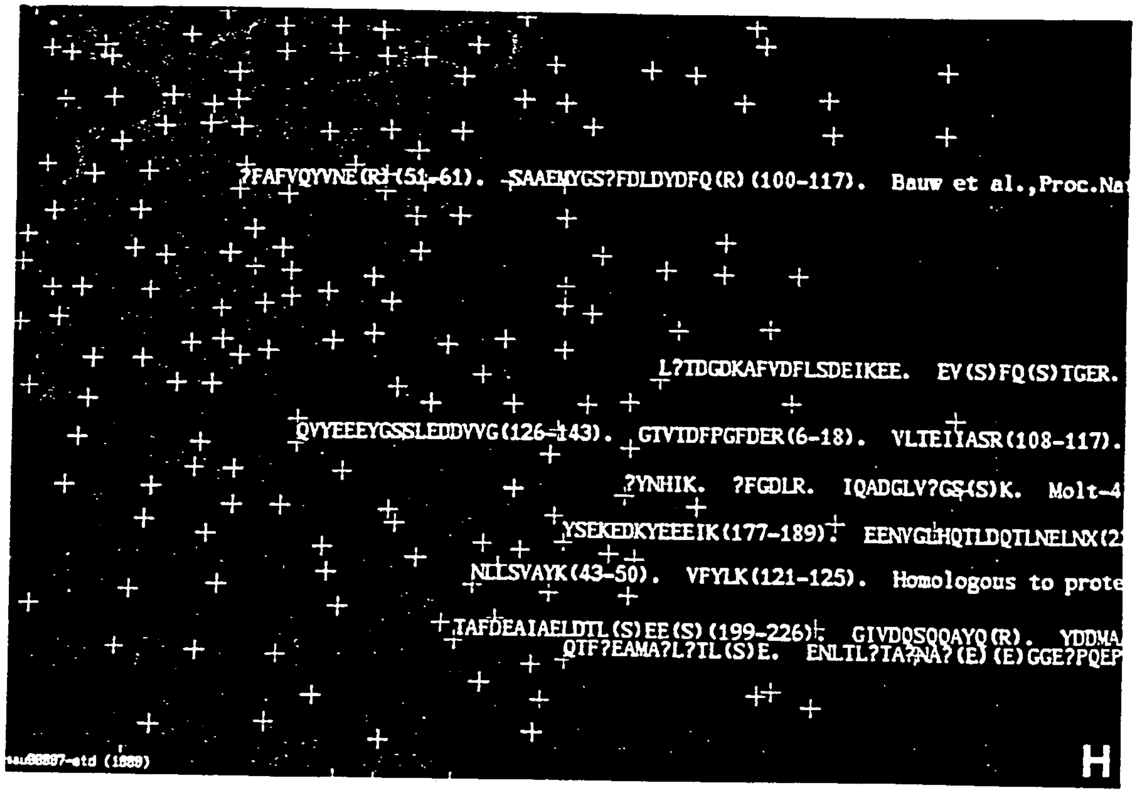


Figure 2. *A*) Synthetic image of a fraction of an IEF gel of the master image of AMA cellular proteins. *B*) As in *A* but showing numbers assigned to each spot. *C*) Comparison of AMA (left) and normal human embryonal lung MRC-5 fibroblasts (right) IEF proteins patterns. Matched proteins are indicated by a + or by the same letters in both gels. Once a protein is matched, information contained in the various categories available in the master AMA database can be transferred. *D*) Synthetic image of a fraction of an IEF fluorogram of [^{35}S]methionine labeled proteins from normal human MRC-5 fibroblasts. The histograms show levels of synthesis of a few proteins in MRC-5 (left bar) and SV40 transformed MRC-5 (right bar) fibroblasts. *E*) Polypeptides that contain information under the category glycolytic pathway. *F*) The function peruse annotation for spot allows the operator to inquire about categories and information available for a given protein. *G*) Relative abundance of cytoskeletal and cytoskeletal-related proteins in quiescent, proliferating, and SV40-transformed MRC-5 fibroblasts. *H*) Polypeptides that contain information under the category partial amino acid sequences.



G



The automatic matching process that has been described in detail by Garrels et al. (12) takes about 5 min. Matched proteins are indicated with the same letters in both gels (Fig. 2C). The usefulness of this function is emphasized by the fact that data accumulated on common household proteins can be easily transferred to any other human cellular cell type whose 2-dimensional gel cellular protein pattern is matched to our standard AMA 2-dimensional gel protein image. Alternatively, if the standard gel is part of a matchset (set of gels in a given experiment) it can be used as a linker gel to compare, for example, the quantitative values of a given protein throughout the experiment (see Fig. 2*D*; levels of some proteins in normal and SV40 transformed human MRC-5 fibroblasts) or with other standard images in different sets of

cross-matched experiments (18, 22).

Once a standard map of a given protein sample is made, one can enter qualitative annotations to make a reference database. Our master 2-dimensional gel database of transformed human amnion cell (AMA) proteins (20) lists 3430 polypeptides of which 2592 correspond to cellular components, having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kDa. The most abundant proteins in the database correspond to total actin (3.87% of total protein; about 90 million molecules per cell) while the lesser abundant of the recorded polypeptides are present in the vicinity of 5000 molecules per cell. Some annotation categories we are using to establish the master AMA database include: *1*) protein identification (comigration with purified proteins, 2-dimensional immunoblotting, microsequencing); *2*) amounts (total amounts and levels of synthesis); *3*) subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); *4*) antibodies; *5*) posttranslational modifications (phosphorylation, glycosylation, methylation etc.); *6*) microsequencing; *7*) cell cycle specificity (specific variations in levels of synthesis and amount); *8*) regulatory behavior (effect of hormones, growth factors, heat shock, etc.) *9*) rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, oncogenes, components of the pathway (or pathways) that control cell proliferation); *10*) function (mainly from comigration with proteins of known function); *11*) sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various cells, etc.); *12*) cDNAs (cloned cDNAs); *13*) proteins that are specific to a given disease (systematic comparison of protein patterns of fibroblast proteins from healthy and diseased individuals); *14*) expression and exploitation of transfected cDNAs; *15*) pathways (metabolic, others); *16*) gene localization (genetic and physical); *17*) effect of microinjected antibody on patterns of protein synthesis; and *18*) secreted proteins.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the color screen. For example, Fig. 2*E* shows a synthetic image of a NEPHGE gel (master AMA database) displaying the information contained under the entry glycolytic pathway. Alternatively, one can use the function peruse annotations for spot to directly ask the computer to list all the entries available for a particular protein. By clicking the mouse in a given entry (in this case, presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Fig. 2*F*).

A major obstacle encountered in building comprehensive 2-dimensional gel protein databases is identifying the large number of proteins separated by this technology. In our databases (20, 21), known proteins are identified by one or a combination of the following procedures: *1*) comigration with known proteins, *2*) 2-dimensional gel immunoblotting using specific antibodies, and *3*) microsequencing of Coomassie Brillant Blue stained human proteins recovered from dried 2-dimensional gels (see next section). Protein identification by means of microsequencing may be difficult, as individual protein members of families with short peptide differences may escape detection. In the gene-protein database of *E. coli* K-12 (14, 23), another major 2-dimensional gel database available at present, proteins are being identified by a wider range of tests that include comigration with purified proteins; genetic criterion (deletion, insertion, frameshift, nonsense, missense, regulatory), plasmid-bearing strains and in vitro synthesis of protein; selective labeling (methylation, phosphorylation); peptide map similarity; and physiological criterion and selective derivatization.

So far we have received nearly 550 antibodies from laboratories all over the world and these are being systematically tested by 2-dimensional gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided identification of unknown proteins (20, 21). We routinely request antibodies and protein samples and promise the donors to make available all the information we may have accumulated on that particular protein. For example, **Table 1** lists entries available for Lipocortin V (IEF SSP 8216), also known as annexin V, VAC-α, endonexin II, renocortin, chromobindin-5', anticoagulant protein, PAP-I, γ-calcimedin, IBC, calphobindin, and anchorin CII.

As mentioned previously, one distinct advantage of 2-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far using computer analysis have been carried out by Garrels and co-workers (18, 22). In particular, these investigators have established a quantitative rat protein database (18, 22) designed to study growth control (proliferation, growth inhibitors, and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus, and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of coregulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (45).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (20, 21). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult without the assistance of a master reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [^{35}S]methionine-labeled proteins synthesized by quiescent, proliferating, and SV40 transformed human embryonic lung MRC-5 fibroblasts (21). Some data concerning cytoskeletal and cytoskeletal-related proteins are presented in Fig. 2G. Our studies as well as those of Garrels and co-workers (18, 22) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

OTHER 2-DIMENSIONAL GEL PROTEIN DATABASES

As mentioned previously there are other 2-dimensional gel databases available in computer form that have been pub-

TABLE 1. *Some entries for lipocortin V in the human AMA 2-dimensional gel protein database*

Entries for lipocortin V (IEF SSP 8216)	Information entered
1. Protein name	Lipocortin V, renocortin, chromobindin-5', endonexin I, anticoagulant protein, PAP-I, VAC-α, 35-γ-calcimedin, IBC, calphobindin I, anchorin CII, annexin V
2. Percentage of total protein	0.110% (about 2,800,000 molecules per cell)
3. Apparent molecular weight (*mr*)	33.3 kDa
4. Isoelectric point (pI)	4.76
5. Method (or methods) of identification	Microsequencing, 2-dimensional immunoblotting, Comigration
6. Credit to investigators that aided in identification	G. Bauw, J. Vandekerckhove, and colleagues, Rijksuniversiteit Gent; B. Pepinsky, BIOGEN, Cambridge; N.G. Ahn, University of Washington
7. Antibody against protein	Polyclonal (rabbit, antibody no. 20), B. Pepinsky, BIOGEN, Cambridge
8. Comigration with human proteins	Lipocortin V,N.G. Ahn, Howard Hughes Medical Institute, Washington University
9. Cellular localization	Subcortical membrane
10. Calcium/phospholipid-dependent membrane proteins	Lipocortin V
11. Function	Regulation of various aspects of inflammation, immune response, blood coagulation and differentiation
12. Partial amino acid sequence	GTVTDFPGFDER (7-18), VLTEIIASR (109-117), QVYEEEYGSSLEDDVVG (127-143), ?GTDEEKFITIFGT(R) (187-201)
13. cDNA sequence	Known, R. Blake et al., *J. Biol. Chem.* 263, 10799-10811; 1988 (pI = 4.76 from translated sequence)
14. Levels in fetal human tissues	Adrenal glands = + + +; brain = + + +; cerebellum = + + +; ear = + + +; eye = + + +; heart = + + +; hypophysis = + + +; liver = + + +; lung = + + +; meninges = + + +; mesonephric tissue = + + +; striated muscle = + + +; pancreas = + + +; skin = + + +; spleen = + + +; stomach = + + +; submandibular gland = + + +; small intestine = + + +; thymus = + + +; thyroid gland = + + +; tongue = + + +; ureter = + + +
15. Levels in quiescent, proliferating, and transformed MRC-5 fibroblasts	Q (quiescent) = 1.1; P (proliferating) = 1.0; T (SV40 transformed) = 0.3
16. Distribution in Triton supernatant and cytoskeletons	Mainly supernatant

lished in extenso: these correspond to the *E. coli* K-12 protein-gene database (14, 23) and to the rat REF52 database (18, 22).

The *E. coli* K-12 cellular protein-gene database is perhaps the most complete of all databases reported so far and eventually it should trace each protein back to its structural gene. Information contained in this database includes: gene/protein name (protein name, EC number, gene name); 2-dimensional gel spot designations (x-y coordinates from reference gels, alphanumeric designation); genetic information (linkage map location, physical map location, Genebank code, sequence reference, location on Kohara clones); biochemical information (molecular weight, pI, number of residues of each amino acid, mole percent of each amino acid, total number of amino acids in a polypeptide), and regulatory information (cellular level of protein in different media and different temperature, member of regulon, member of stimulon). Major advances of this database are envisaged in the future in view of the eminent sequencing of the whole *E. coli* genome as well as the development of improved methods to express cloned genes.

The rat REF52 2-dimensional gel protein database lists about 1600 proteins that have been recorded using the QUEST analysis system (18, 22). Included in this quantitative database are *1*) protein names (cytoskeletal and heat shock proteins as well as various nuclear, mitochondrial, and cytoplasmic proteins), *2*) annotations (subcellular localization, modification, recognition by specific antibodies, coprecipitation, NH_2-terminal sequence, cross-reference to protein sequence information and references to the literature), *3*) protein sets (cytoskeletal proteins, phosphoproteins, sets of proteins with PCNA/cyclin-like properties, etc.) and *4*) general quantitative data (protein synthesis during growth of normal REF52 cells to confluence and quiescence, and after restimulation of growth-inhibited cells).

In addition to the 2-dimensional gel databases mentioned so far there are several smaller cellular databases being established in human (normal human diploid fibroblasts, lym-

phocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T lymphocytes), *Aplysia*, yeast (*Saccharomyces cerevisae*), plants (wheat, barley, sorghum), and *Euglena*. Databases of tissue protein, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine, and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (4) for details and references concerning these databases.

MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE 2-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly sensitive amino acid gas-phase or liquid-phase sequenators (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a systematic sequence analysis of proteins resolved by 2-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks) as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Fig. 1). In addition, partial protein sequences can be stored in 2-dimensional gel databases (for example, see Fig. 2*H*) and offer a unique link between proteins and genes (Fig. 1).

In the early 1970s gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in ref 25). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (26). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and NH_2-terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the 10 picomol level).

Most of the problems referred to above have been minimized with the introduction of protein-electroblotting procedures (27–32). When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields NH_2-terminal sequences containing 10–40 residues. As such, this technique was used to identify, by their NH_2-terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2-dimensional gels. A major difficulty encountered in this procedure is the occurrence of frequent artefactual blockage of the proteins. Several studies suggest that this phenomenon is mainly due to reaction with contaminants (particularly unpolymerized acrylamide present in the gel) and to a high dilution of the protein (low concentration of the protein per unit membrane surface). In addition to this primarily technical problem, many proteins are blocked in vivo by acylation or by a pyrrolidon carboxylic acid cap.

The problem of partial or complete NH_2-terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel in situ cleavage) or by cleaving it while bound to the membrane (membrane in situ cleavage) (33–35). In both cases, proteins are either cleaved in a restricted way (e.g., by limited enzymatic digestion or by using restriction chemical cleavage conditions) or fragmented into smaller peptides.

Of the different combinations examined, we had good results by using exhaustive proteolytic digestion of membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots (34), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass fiber membranes (35). The proteases used (trypsin, chymotrypsin, or pepsin) cleave at multiple sites, generating small peptides that elute from the blot into the digestion buffer from which they are purified by reversed-phase high performance liquid chromatography (HPLC) before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we were surprised that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel-purified proteins available in amounts ranging from 5 to 10 μg, and often yielded sequence information covering more than 30% of the total protein. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein database (20) (a few are shown in Fig. 2*H*) have been obtained from analysis of protein spots collected from 2-dimensional gels that had been stained with Coomassie blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by a protein-elution-concentration device, combined with gel electrophoresis and electroblotting. Details of this technique have been reported in a previous communication (42) and a brief outline is given below.

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then collected into a large slot made in a new gel. The slot is further filled with Sephadex G-10 equilibrated in gel sample buffer. During consecutive gel electrophoresis, most of the electrical current passes on the side of the slot instead of passing through the slot. This results in both a vertical stacking and horizontal contraction of the protein band. With this device the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot. The highly concentrated (about 5 mm^2) protein spot is then electroblotted on PVDF-membranes, stained with Amido black, and in situ digested with trypsin. The peptides generated during digestion elute from the membrane into the supernatant, and can be separated by narrow bore reversed-phase HPLC and collected individually for sequence analysis.

Using this and previous procedures (37, 39, 42), we have so far analyzed 70 protein spots collected from 2-dimensional gels (20, and unpublished observations) (see for example Fig. 2*H*). The sequence information amounts to 2100 allocated residues corresponding to an average of 30 residues per protein spot. So far we have made cDNAs of many of the unknown proteins that have been microsequenced, and a substantial number has been cloned and sequenced. All available information indicates that it may be possible to obtain partial sequence information from most of

the proteins that can be visualized by Coomassie Brillant Blue staining.

Partial protein sequences are stored in the database as displayed in Fig. 2*H*. and it should be possible in the near future to interface this information with forthcoming DNA sequence data from the human genome project. In the long run. as the human genome sequences become available it will be possible to assign partial protein sequences to genes for which the full DNA sequence and chromosomal location are known (Fig. 1).

SUMMARY

The studies presented in this brief review are intended to demonstrate the usefulness of computer-aided 2-dimensional gel electrophoresis and microsequencing to analyze cellular protein patterns. and to link protein and DNA information. As more information is gathered worldwide. comprehensive databases will depict an integrated picture of the expression levels and properties of the thousands of proteins that orchestrate most cellular functions.

Clearly. databases allow easy access to a large body of data and provide an efficient medium to communicate standardized protein information. In the future. databases will foster a wide variety of biological information that can be used to support collaborative research projects in basic and applied biology as well as in clinical research (2. 5. 46). Once a protein is identified in a particular database all the information gathered on it can be made available to the scientist. However. many problems must be solved before protein databases become of general use to the scientific community. A most urgent one is to promote standardization of the gel running conditions so that data produced in a given laboratory may be used worldwide. Surprisingly. the gel running technology as it stands today is still a craftmanship art.

Finally, comprehensive, computerized databases of proteins. together with recently developed techniques to microsequence proteins. offer a new dimension to the study of genome organization and function (Fig. 1). In particular. human protein databases may become increasingly important in view of the concerted effort to map and sequence the entire human genome. This formidable task is expected to dominate biological research in the next decades. FJ

We would like to thank S. Himmelstrup Jørgensen for typing the manuscript and O. Sønderskov for photography. Work in the authors' laboratories was supported by grants from the Danish Biotechnology Programme. the Danish Cancer Foundation. and the Commission of the European Communities.

REFERENCES

1. O'Farrell. P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**. 4007–4021
2. Special Issue: Two-dimensional gel electrophoresis. *Clin. Chem.* **28**. 1982
3. Celis. J. E.. and Bravo. R.. eds. (1984) *Two-Dimensional Gel Electrophoreis of Proteins: Methods and Applications.* Academic. New York
4. Celis. J. E.. Madsen. P.. Gesser. B.. Kwee. S.. Nielsen. H. V.. Rasmussen. H. H.. Honoré. B.. Leffers. H.. Ratz. G. P.. Basse. B.. Lauridsen. J. B.. and Celis. A. (1989) Protein databases derived from the analysis of two-dimensional gels. In *Advances in Electrophoresis* (Chrambach. C.. ed) VCH. Weinheim. Germany
5. Special Issue: Two-dimensional gel electrophoresis in cell biology. (Celis. J. E.. ed) *Electrophoresis* **11**. 1990
6. Celis. J. E.. Honoré. B.. Bauw. G.. and Vandekerckhove. J (1990) Comprehensive computerized 2D gel protein databases offer a global approach to the study of the mammalian cell *BioEssays* **12**. 93–98
7. Garrels. J. I. (1983) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by cloned cell lines *Methods Enzymol.* **100**. 411–423
8. Anderson. N. L.. Hofmann. J. P.. Gemmel. A.. and Taylor. S. (1984) Global approaches to the quantitative analysis of gene-expression patterns observed by two-dimensional gel electrophoresis. *Clin. Chem.* **30**. 2031-2036
9. Garrels. J. I.. Farrar. J. T.. and Burwell. C. B. (1984) The Quest system for computer-analyzed two-dimensional electrophoresis of proteins in *Two-Dimensional Gel Electrophoresis of Proteins. Methods and Applications* (Celis. J. E.. and Bravo. R.. eds) pp. 37–91. Academic. New York
10. Vincens. P.. and Tarroux. P. (1988) Two-dimensional electrophoresis computerized processing. *Int. J. Biochem.* **20**. 499–509
11. Appel. R.. Hochstrasser. D.. Roch. C.. Funk. M.. Muller. A. F.. and Pellegrini. C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **9**. 136–142
12. Lemkin. P. F.. and Lester. E. P. (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* **10**. 122–140
13. Miller. M. J. (1989) Computer-assisted analysis of two-dimensional gel electrophoretograms. *Adv. Electrophoresis* **3**. 182–217
14. Phillips. T. D., Vaughn. V., Bloch. P. L.. and Neidhardt. F. C. (1987) In *Eschericia coli and Salmonella typhimurium: Cellular and Molecular Biology, Gene-Protein Index of Escherichia coli K-12*, 2 ed. (Neidhardt. F. C.. Ingraham. J. I.. Low. K. B.. Magasanik. B.. Schaechter. M.. and Umbarger. H. E. ed) pp. 919–966. American Society for Microbiology. Washington. D.C.
15. Celis. J. E.. Ratz. G. P.. Celis. A.. Madsen. P.. Gesser. B.. Kwee. S.. Madsen. P. S.. Nielsen. H. V.. Yde. H.. Lauridsen. J. B.. and Basse. B. (1988) Towards establishing comprehensive databases of cellular proteins from transformed human epithelial amnion cells (AMA) and normal peripheral blood mononuclear cells. *Leukemia* 9. 561–601
16. Special Issue: Protein databases in two-dimensional electrophoresis. (Celis. J. E.. ed) *Electrophoresis* 2. 1989
17. Celis. J. E.. Ratz. G. P.. Madsen. P.. Gesser. B.. Lauridsen. J. B.. Brogaard-Hansen. K. P.. Kwee. S.. Rasmussen. H. H.. Nielsen. H. V.. Crüger. D.. Basse. B.. Leffers. H.. Honoré. B.. Møller. O.. and Celis. A. (1989) Computerized. comprehensive databases of cellular and secreted proteins from normal human embryonic lung MRC-5 fibroblasts: identification of transformation and/or proliferation sensitive proteins. *Electrophoresis* **10**. 76-115
18. Garrels. J. I.. and Franza. B. R. (1989) The REF52 protein database. Methods of database construction and analysis using the Quest system and characterizations of protein patterns from proliferating and quiescent REF52 cells. *J. Biol. Chem.* **264**. 5283–5298
19. Celis. J. E.. Crüger. D.. Kiil. J.. Dejgaard. K.. Lauridsen. J. B.. Ratz. G. P.. Basse. B.. Celis. A.. Rasmussen. H. H.. Bauw. G.. and Vandekerckhove. J. (1990) A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis* **11**. 242–254
20. Celis. J. E.. Gesser. B.. Rasmussen. H. H.. Madsen. P.. Leffers. H.. Dejgaard. K.. Honoré. B.. Olsen. E.. Ratz. G.. Lauridsen. J. B.. Basse. B.. Mouritzen. S.. Hellerup. M.. Andersen. A.. Walbum. E.. Celis. A.. Bauw. G.. Puype. M.. Van Damme. J.. and Vandekerckhove. J. (1990) Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* **12**. 989-1071

21. Celis. J. E.. Dejgaard. K.. Madsen. P.. Leffers. H.. Gesser. B.. Honoré. B.. Rasmussen. H. H.. Olsen. E.. Lauridsen. J. B.. Ratz. G.. Mouritzen. S.. Hellerup. M.. Andersen. A.. Walbum. E.. Celis. A.. Bauw. G.. Puype. M.. Van Damme. J.. and Vandekerckhove. J. (1990) The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent. proliferating and SV40 transformed cells. *Electrophoresis* **12**. 1072–1113

22. Garrels. J. I.. Franza. B. R.. Chang. C.. and Latter. G. (1990) Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation. serum stimulation. and viral transformation. *Electrophoresis* **12**. 1114–1130

23. Van Bogelen. R. A.. Hutton. M. E.. and Neidhardt. F. C. (1990) Gene-protein database of *Escherichia coli* K-12. 3rd ed. *Electrophoresis* **12**. 1131–1166

24. Hewick. R. M.. Hunkapiller. M. W.. Hood. L. E.. and Dreyer. W. J. (1981) A gas-liquid solid phase peptide and protein sequenator. *J. Biol. Chem.* **256**. 7990–7997

25. Weber. K.. and Osborn. M. (1985) In *The Proteins and Sodium Dodecyl Sulfate: Molecular Weight Determination on Polyacrylamide Gels and Related Procedures* (Neurath. H. et al.. eds) Vol. 1. pp. 179–223. Academic. New York

26. Hunkapiller. M. W.. Lujan. E.. Ostrander. F.. and Hood. L. E. (1983) Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. *Methods Enzymol.* **91**. 227–236

27. Vandekerckhove. J.. Bauw. G.. Puype. M.. Van Damme. J.. and Van Montagu. M. (1985) Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* **152**. 9–19

28. Aebersold. R. H.. Teplow. D. B.. Hood. L. E.. and Kent. S. B. H. (1986) Electroblotting onto activated glass. *J. Biol. Chem.* **261**. 4229–4238

29. Bauw. G.. De Loose. M.. Inzé. D.. Van Montagu. M.. and Vandekerckhove. J. (1987) Alterations in the phenotype of plant cells studied by NH_2-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. *Proc. Natl. Acad. Sci. USA* **84**. 4806–4810

30. Matsudaira. P. (1987) Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *J. Biol. Chem.* **262**. 10035–10038

31. Eckerskorn. C.. Mewes. W.. Goretzki. H.. and Lottspeich. F. (1985) A new siliconized-glass fiber as support for protein-chemical analysis of electroblotted proteins. *Eur. J. Biochem.* **176**. 509–519

32. Moos. M.. Jr.. Nguyen. N. Y.. and Liu. T-Y. (1988) Reproducible high yield sequencing of proteins electrophoretically separated and transferred to an inert support. *J. Biol. Chem.* **263**. 6005–6008

33. Kennedy. T. E.. Gawinowicz. M. A.. Barzilai. A.. Kandel. E. R.. and Sweatt. J. D. (1988) Sequencing of proteins from two-dimensional gels by using in situ digestion and transfer of peptides to polyvinylidene difluoride membranes: application to protein associated with sensitization in *Aplysia*. *Proc. Natl. Acad. Sci. USA* **85**. 7008–7012

34. Aebersold. R. H.. Leavitt. J.. Saavedra. R. A.. Hood. L. E.. and Kent. S. B. H. (1987) Internal amino acid sequence analysis of protein separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* **84**. 6970–6974

35. Bauw. G.: Van Den Bulcke. M.. Van Damme. J.. Puype. M. Van Montagu. M.. and Vandekerckhove. J. (1988) Protein electroblotting on polybase-coated glassfiber and polyvinylidine difluoride membranes: an evaluation. *J. Prot. Chem.* **7**. 194–196

36. Celis. J. E.. Ratz. G. P.. Madsen. P.. Gesser. B.. Lauridsen. J. B.. Leffers. H.. Rasmussen. H. H.. Nielsen. H. V.. Crüger. D.. Basse. B.. Honoré. B.. Moller. O.. Celis. A.. Vandekerckhove. J.. Bauw. G.. Van Damme. J.. Puype. M.. and Van Den Bulcke. M. (1989) Comprehensive. human cellular protein databases and their implication for the study of genome organization and function. *FEBS Lett.* **244**. 247–254

37. Bauw. G.. Van Damme. J.. Puype. M.. Vandekerckhove. J.. Gesser. B.. Lauridsen. J. B.. Ratz. G. P.. and Celis. J. E. (1989) Protein-electroblotting and -microsequencing strategies in generating protein databases from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* **86**. 7701–7705

38. Aebersold. R.. and Leavitt. J. (1990) Sequence analysis of proteins separated by polyacrylamide gel electrophoresis. Towards an integrated protein database. *Electrophoresis* **11**. 517–527

39. Bauw. G.. Rasmussen. H. H.. Van Den Bulcke. M.. Van Damme. J.. Puype. M.. Gesser. B.. Celis. J. E.. and Vandekerckhove. J. (1990) Two-dimensional gel electrophoresis. protein electroblotting and microsequencing: a direct link between proteins and genes. *Electrophoresis* **11**. 528–536

40. Tempst. P.. Link. A. J.. Riviere. L. R.. Fleming. M.. and Elicone. C. (1990) Internal sequence analysis of protein separated on polyacrylamide gels at the submicrogram level: improved methods. applications and gene cloning strategies. *Electrophoresis* **11**. 537–553

41. Eckerskorn. C.. and Lotspeich. F. (1990) Combination of two-dimensional gel electrophoresis with microsequence and amino acid composition analysis: improvement of speed and sensitivity in protein characterization. *Electrophoresis* **11**. 554–561

42. Rasmussen. H. H.. Van Damme. J.. Bauw. G.. Puype. M.. Gesser. B.. Celis. J. E.. and Vandekerckhove. J. (1991) In *Methods in Protein Sequence Analysis* (Jörnvall. H.. and Höög. J. O.. eds) pp. 103–114. Eighth International Conference on Methods in Protein Sequence Analysis. Birkhäuser Verlag. Boston

43. Olson. A. D.. and Miller. M. J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* **169**. 49–70

44. Vincens. P.. Paris. N.. Pujol. J. L.. Gaboriaud. C.. Rabilloud. T.. Pennetier. J.. Matherat. P.. and Tarroux. P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* **7**. 347–356

45. Celis. J. E.. Madsen. P.. Celis. A.. Nielsen. H. V.. and Gesser. B. (1987) Cyclin (PCNA. auxiliary protein of DNA polymerase-δ) is a central component of the pathway(s) leading to DNA replication and cell division. *FEBS Lett.* **220**. 1–7

46. Anderson. N. G.. and Anderson. N. L. (1982) The human protein index. *Clin. Chem.* **28**. 739–748

Electrophoresis 1993, *14*, 1045–1053

Bo Franzén[1]
Stig Linder[3]
Ken Okuzawa[2]
Harabumi Kato[2]
Gert Auer[1]

[1]Division of Tumor Pathology, Department of Pathology, Division of Experimental Oncology, Karolinska Hospital and Institute, Stockholm Sweden
[2]Tokyo Medical College, Department of Surgery, Tokyo
[3]Division of Experimental Oncology, Karolinska Hospital and Institute, Stockholm

Nonenzymatic extraction of cells from clinical tumor material for analysis of gene expression by two-dimensional polyacrylamide gel electrophoresis

We have compared different methods of preparation of malignant cells for two-dimensional electrophoresis (2-DE). We found all methods using fresh tissue to be superior compared to methods using frozen tissue. Our results indicate that nonenzymatic methods of preparation of tumor cells, including fine needle aspiration, scraping and squeezing, have advantages over methods using enzymatic extraction of cells. Nonenzymatic methods are rapid, appear to reduce loss of high molecular protein species, and alleviate the necessity of separating viable and nonviable cells by Percoll gradient centrifugation. Using these techniques, high-quality 2-DE maps were derived from tumors of the lung and breast. In the resulting polypeptide patterns, heat shock proteins, non-muscle tropomyosins and intermediate filament were identified. We conclude that nonenzymatic extraction of malignant cells from fresh tumor tissue improves the possibilities that these techniques may be useful in clinical diagnosis.

1 Introduction

Tumors may develop by a number of different mechanisms in any given cell type. At the time of diagnosis, tumors will have progressed along different pathways to various stages of malignancy. To provide a basis for individual therapy it is of importance to examine specific properties of the tumor cell population in each patient. A large number of different markers have been described in order to increase the diagnostic accuracy. It is likely that a combination of serveral markers is needed in the future in order to reflect different properties of the tumor. One important method for the resolution of a large number of potential markers is two-dimensional electrophoresis (2-DE). Extensive efforts are being made in identifying various polypeptides separated by 2-DE and to characterize how the expression of these polypeptides is affected by the response to cellular transformation and various culture conditions [1,2]. It would be of value to transfer this information to 2-DE separations of polypeptides from tumor tissue samples. However, one prerequisite is that the quality of the 2-DE gels from tumor samples is comparable in quality with 2-DE gels from samples of cultured cells.

Frozen tumor tissues are commonly used for various biochemical assessments. However, if such samples are analyzed by 2-D polyacrylamide gel electrophoresis (PAGE), the polypeptide patterns are obscured by contamination of serum- and connective tissue proteins. Such nontumor-cell-related variations represent serious problems in the interpretation and inter-patient comparison of 2-DE patterns [3]. 2-DE patterns of cells prepared from fresh tumor material were analyzed after enzymatic extraction of tumor cells [4, 5] or after culturing tumor fragments in medium containing radioactive amino acids [6]. These procedures may, however, lead to alterations in the gene expression/polypeptide patterns. We are only aware of one study where nonenzymatic extraction of cells from fresh tumor tissue (prostate cancer) was used to prepare samples for 2-D PAGE [4]. We have examined enzymatic extraction and various nonenzymatic preparation techniques, including fine needle aspiration, for the preparation of cells from fresh tumor tissues. We describe nonenzymatic extraction procedures that are rapid, lead to high-quality 2-DE patterns, and that alleviate the necessity to purify tumor cell populations from dead cells.

2 Materials and methods

2.1 Cell cultures and samples used for spot identification

A rat embryonal fibroblast cell line, WT2 (a kind gift from Dr. J. I. Garrels and Dr. S. Pattersson) was used for the identification of a number of heat shock and structural proteins. Human normal diploid lung fibroblasts, WI38, human epithelial breast carcinoma cells, MDA-231 and MCF-7 were purchased from ATCC and grown as recommended. Polypeptides prepared from a leukemia type pre-B-ALL were separated by 2-DE. The 2-DE map was then analyzed by Dr. S. M. Hanash (University of Michigan, Ann Arbor, USA).

2.2 Tumor tissues samples

In this study, 2-DE maps from seven tumors were used as representative illustrations: two adenocarcinoma of the lung (LA, and LB, mucinous, both cases intermediate grade of differentiation), one sqamous carcinoma of the lung (LS), one carcinoid-like breast cancer (BC), one microfolliculary adenoma (highly differentiated) of the thyroid (TA), one highly differentiated hyperneph-

Correspondence: Dr. Bo Franzen, Division of Tumor Pathology, Department of Pathology, L1:01, Karolinska Hospital and Institute, 10401 Stockholm 60, Sweden

Abbreviations: **2-DE**, Two-dimensional polyacrylamide gel electrophoresis; **IEF**, isoelectric focusing; **LDH**, lactate dehydrogenase; **NP-40**, Nonidet P-40; **PBS**, phosphate buffered saline; **PCNA**, proliferating cell nuclear antigen; **PIH**, protease inhibitors; **PMSF**, phenylmethyl sulfonyl fluoride; **SDS**, sodium dodecyl sulfate; **WW**, wet weight

0173-0835/93/1010-1045 $5.00+.25/0

roma, a tumor of the kidney (KH), and finally one case of poorly differentiated corpus carcinoma (CP).

2.3 Preparation of cultured cells

The cell monolayers were washed twice in phosphate buffered saline (PBS) and then scraped off in ice-cold PBS including protease inhibitors (PIH), phenylmethylsulfonyl fluoride (PMSF) 0.2 mM and 0.83 mM benzamidine pelleted at 660 × g, 3 min (+4°C) and washed one time before final centrifugation at 2700 × g, 5 min. The wet weight of the cell pellet was recorded and the cells were stored at −80°C until further processing.

2.4 Preparation of tumor tissue samples

2.4.1 General remarks

Macroscopically representative and non-necrotic tumor tissues were selected within 20 min after resection. Parallel samples were routinely prepared for cytology. The samples were processed as rapidly as possible on ice or at +4°C and in the presence of PIH. Cells were stained with DiffQuick (Baxter) and usually examined at three different occasions during the preparation procedure: (i) cytology sample, (ii) extracted cells and (iii) cells after percoll gradient centrifugation.

2.4.2 Specimen acquisition

The strategy of sample preparation is shown in Fig. 1. Tumor tissue cell samples were usually obtained by fine needle aspiration (NA) using a 0.7 mm needle. The syringe was filled with 1–2 mL of ice-cold culture medium/PIH. We found that if a tumor appeared to be very fibrous it is difficult to extract enough cells for 2-DE analysis. In these cases, two alternative techniques were examined. (i) The tumor was cut in the middle and the fresh surface scraped (SC) by a scalpel. The cell-rich material was then transferred to ice-cold culture medium (L15 with 5% fetal calf serum)/PIH. (ii) A part of the tumor sample was placed in culture medium on ice for further processing at the laboratory in the following way: the material was cut into very small fragments on a pre-cooled dissection plate and transferred to a small glass chamber with a 0.7 mm metal net 5 mm above the bottom of the chamber. Medium /PIH was added to cover the sample (8 mL) which was gently squeezed (SQ) towards the net in order to release and wash out cells. NA and SC were also compared with an enzymatic extraction (EE) procedure described previously [5]: Briefly, thin slices of tissue were incubated with collagenase (1 mg/mL) and elastase (2 mg/mL) in medium for 1 h at 37°C. Extracted cells from every sample were then subjected to percoll gradient centrifugation (Section 3.2.3).

2.4.3 Separation of cells by Percoll gradient centrifugation

The cell suspension was filtered through two nylon mesh filters, (i) 250 µm and (ii) 100 µm and then centrifuged at 660 × g for 3 min. The cell pellet was resuspended carefully in medium, using a syringe and loaded onto a two-step discontinuous Percoll/PBS gradient, 20.4% (density = 1.03 g/mL) and 54.7% (density = 1.07 g/mL), and centrifuged at 1000 × g for 15 min. In this system, dead cells stay on the top, viable cells sediment to the interphase and erythrocytes sediment to the bottom. The viability of cells in the top fraction and interphase was checked by the trypan blue exclusion test. The interphase cell layer (> 90% viability) was collected and washed one time in a large volume PBS/PIH (centrifuged at 800 × g for 3 min). Finally, the cells were resuspended in 1.4 mL PBS and pelleted at 2700 × g for 5 min. The wet weight (WW) was recorded and the pellet was then stored at −80°C.

2.4.4 Final preparation of cells for 2-D PAGE analysis

From this point, cultured cell samples were treated in the same way as tumor cell samples: Each cell pellet was thawed on ice and resuspended in 1.89 µL mQ water per mg WW (= 1.89 × WW) µL. The suspension was frozen and thawed 4–5 × to break the cells [7]. A volume of (0.089 × WW) µL 10% sodium dodecyl sulfate (SDS), including 33.3% mercaptoethanol, was mixed with the sample and incubated 5 min on ice with (0.329 × WW) µL of a solution of DNase I (0.144 mg/mL 20 mM Tris-HCl with 2 mM CACl$_2$ × 2H$_2$O, pH 8.8) and RNase A (0.0718 mg/mL Tris) [8,9]. The sample was frozen and lyophilized. Sample buffer [10] including

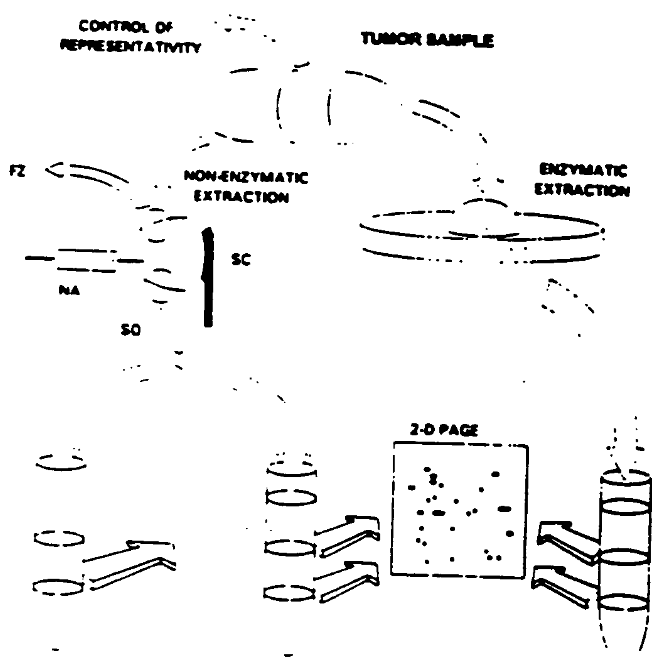


Figure 1. Experimental flow chart showing main steps of the preparation procedures. The abbreviations used for nonenzymatic extraction procedures are: FZ: frozen sample preparation; NA, needle aspiration; SC, scraped; and SQ, squeezed sample. Extracted cells are then loaded as a suspension (top volume of each tube) onto either 1.07 g/mL Percoll (left), or a discontinuous Percoll gradient from the nonenzymatic extraction (middle), or from enzymatic extraction (right). Cellular top- and interphase fractions are then used for 2-DE. For details see Section 2.

PMSF (0.2 mM, EDTA (1.0 mM), 0.5% Nonidet P-40 (NP-40), and 3-[3-cholamido propyl)-dimethylammonio]-1-propane sulfonate (CHAPS; 25 mM) was added carefully, mixed for 2.5 h and centrifuged for 15 min at 10000 rpm to remove any insoluble material. Duplicate or triplicate samples were taken for protein determination [11]. Samples were stored at −80°C prior to isoelectric focusing (IEF).

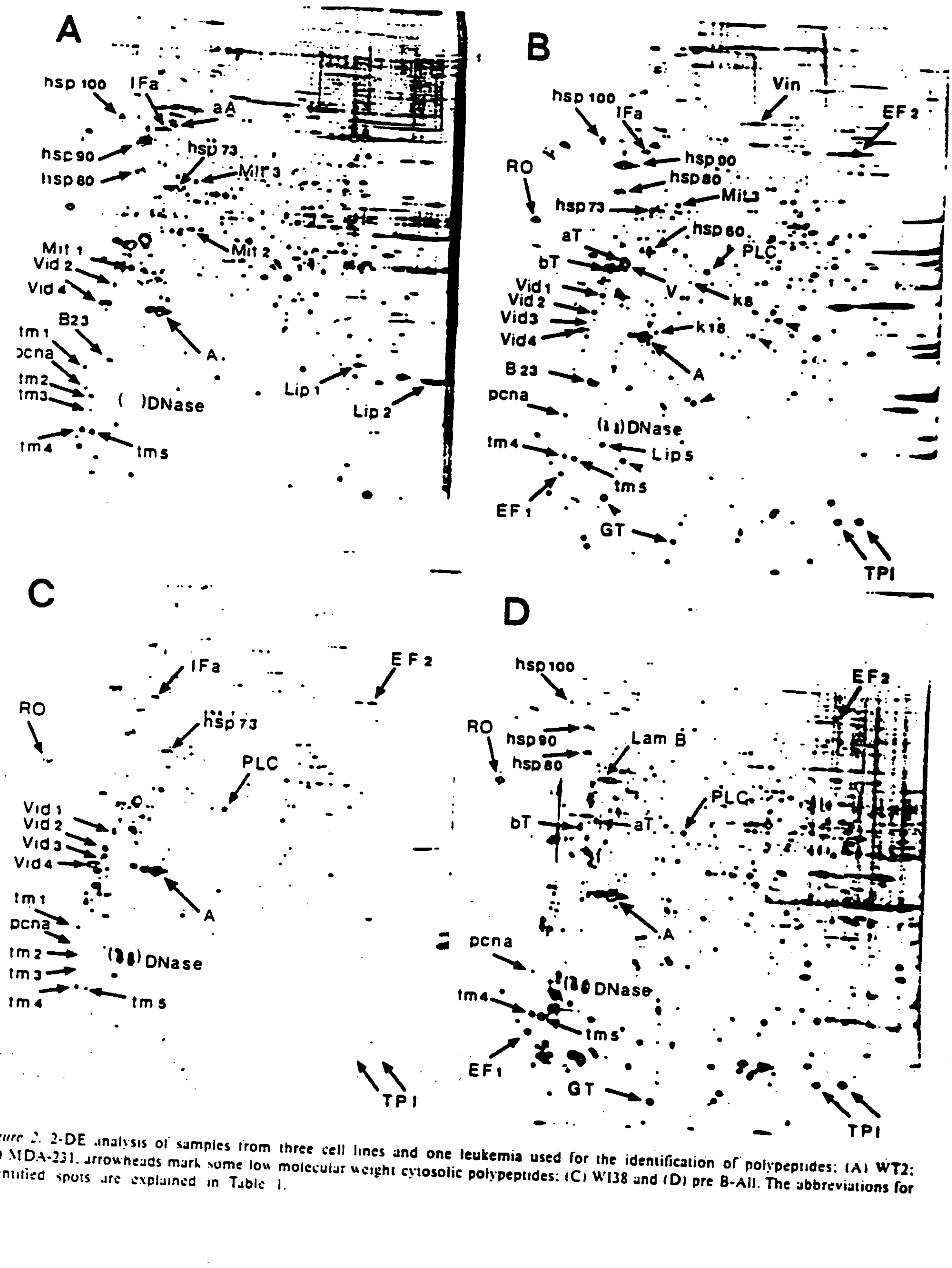


Figure 2. 2-DE analysis of samples from three cell lines and one leukemia used for the identification of polypeptides: (A) WT2; (B) MDA-231, arrowheads mark some low molecular weight cytosolic polypeptides; (C) WI38 and (D) pre B-All. The abbreviations for identified spots are explained in Table 1.

2.4.5 Preparation of frozen tumor tissue

The technique has been described previously [3,12]. Briefly, the sample is moarted frozen to a fine powder, homogenized, lyophilized and solubilized in sample buffer.

2.4.6 Control of representativity

The tumors were examined routinely by experienced pathologists and smears or imprints from the samples were also assessed for cytometric DNA content by microspectrophotometry.

2.5 2-D PAGE

2-D PAGE was performed as described [8,10] except for the following details. The glass tubes for IEF, 1.2 × 200 mm, contained 2.0% Resolyte, pH 4–8 (BDH) and were cast to a height of 180 mm. A stock solution of acrylamide (Serva) and N,N'-methylenebisacrylamide (16.7:1 for IEF and 37.5:1 for the second dimension) was deionized by mixing with 5% w/v Duolite MB 5313 mixed-resin ion exchanger (BDH) for 30 min, filtered (with a 0.22 μm nitrocellulose filter) and stored at −70°C. N,N'-Methylenebisacrylamide, N,N,N',N'-tetramethylethylenediamine (TEMED) and ammonium persulfate were purchased from Bio-Rad. IEF tubes were prefocused at 200 V in 60 min. To each tube a sample corresponding to 20–40 μg protein was applied and focused for 14.5 h at 800 V and finally 1.0 h at 1000 V using a Protean II cell (Bio-Rad) and Model 1000/500 Power Supply (Bio-Rad). The tube gels were finally extruded into 1.25 mL equilibration buffer, containing 60 mM Tris, pH 6.8 (2% SDS, 100 mM dithiothreitol and 10% glycerol), frozen on dry ice and stored at −70°C. The second dimension (1.0 × 180 × 90 mm) of the acrylamide concentration was 10% T, and the gel contained 376 mM Tris, pH 8.8, and 0.1 SDS. IEF gels were applied on top of the slab gel, sealed with 0.5% agarose containing electrophoresis running buffer (60 mM Tris-base, 0.2 M glycine and 0.1% SDS) and electrophoresed with 10–11 mA per gel (constant current) at +10°C. Six gels were run together in a Protean II xi 2-D Multi-Cell (Bio-Rad). Proteins were visualized by silver staining and photographed with the acidic side to the left [13,14].

2.6 Identification of polypeptides

Vimentin and vimentin-derived polypeptides were identified by extraction of an MDA-231 cell lysate with 0.6 M KCl/0.5% NP-40 [15]. Tropomyosins were extracted from MDA-231 and WI38 cell lysates [16], and cytokeratins were extracted from MDA-231 and MCF-7 cell lysates [17]. The patterns were compared with published maps [19–21]. Proliferating cell nuclear antigen (PCNA) was identified by immunoblotting (PC10 mAB, Dakopatt) using a semidry system (Multiphor II Nova Blot, Pharmacia-LKB Biotechnology AB) and enhanced chemoluminescence (ECL) detection (Amersham).

3 Results

3.1 2-DE of samples prepared from normal and tumorigenic cultured cells

The object of this study was to develop methods for preparation of 2-DE maps from human tumor tissue which have the same high resolution as those obtained from cultured cells. Shown in Fig. 2 are high resolution 2-DE gels prepared from cultured cells and one leukemia: SV40 transformed embryonal rat fibroblasts WT2 (Fig. 2a); human MDA-231 breast carcinoma cells (Fig. 2b); human WI38 fibroblasts (Fig. 2c) and human pre B-ALL

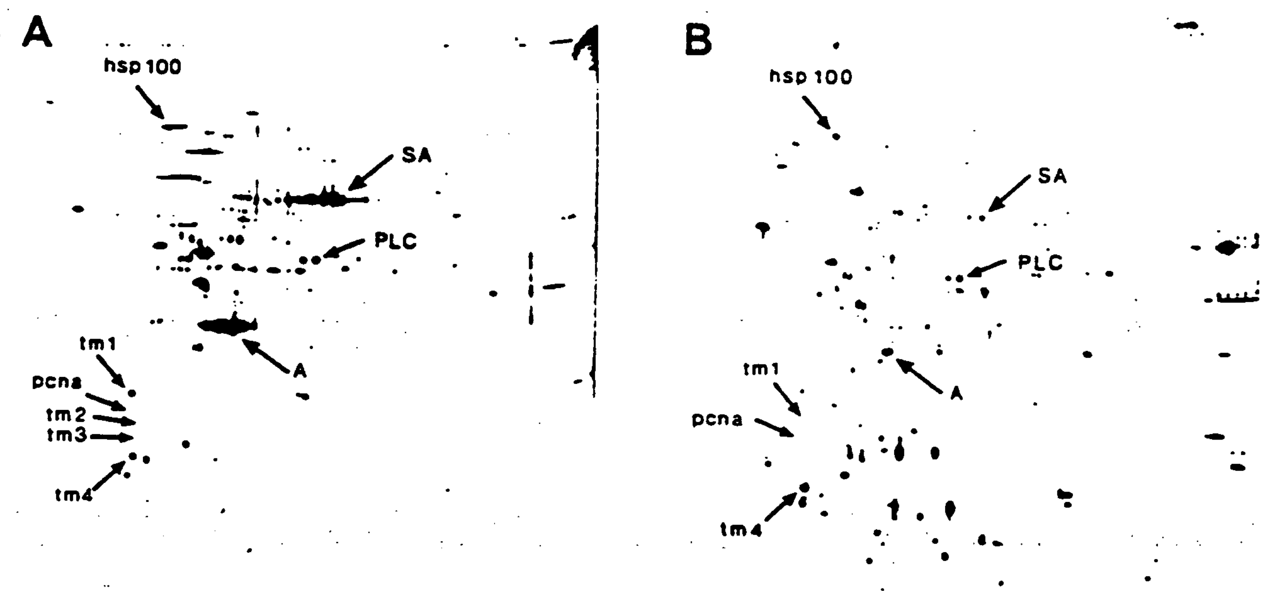


Figure 3. 2-DE analysis of a case of lung adenocarcinoma (LA). Comparison of 2-DE gel quality between (A) frozen and (B) fresh (needle aspiration) tissue preparation.

cells (Fig. 2d). Polypeptides were identified through a laboratory exchange of cell samples/2-DE maps and through 2-DE analysis of purified proteins (Table 1).

3.2 Preparation of samples from solid tumors

3.2.1 Fresh *versus* frozen tissue

An adenocarcinoma of the lung (LA) was prepared for 2-DE by conventional methods using frozen material (Fig. 3a). There are several possibilities for the poor resolution using frozen tissue, including the presence of high molecular weight protein aggregates. Filtering extracts through 0.1 μm filters (Durapore, Millipore) resulted in a slightly improved resolution (not shown). When fresh tumor tissue from tumor LA was used for sample preparation, using fine needle aspiration to collect the cells, the resolution was considerably improved (Fig. 3b). The use of fresh tissue resulted in a general increase in resolution, which was most pronounced in the 50–100 kDa molecular mass range. A number of differences in the protein profiles of the gels in Figs. 3a and 3b can be observed, some of which are indicated in the figures. The decrease in serum albumin in Fig. 3b is likely to result from loss of serum proteins occurring when cells were pelleted after aspiration. Other differences, such as the decreased level of transformation-sensitive tropomyosins (TM1-TM3), may result from enrichment of tumor cells in the sample of Fig. 3b. Fine needle aspiration, a well-established technique in cytology, extracts mainly tumor cells because of decreased intercellular adhesiveness of neoplastic cells as compared to normal tissue. Microscopic examination of Diff-Quick-stained extracted cells from case LA revealed almost 100% tumor cells, whereas the whole tissue extract contained approximately 60% tumor cells.

Table 1. Names and abbreviations for identified spots

Spot	Name	Basis for identification
A	Actins	a
aA	*alpha*-Actinin	a
B23	Protein B23 /Numatrin	a
EF2	Elongation factor 2	a
EF1	Elongation factor 1 β	a
GT	Glutathione-S-transpherase (pi	a
hsp60	Heat shock protein 60	a
hsp73	Heat shock protein 73	a
hsp80	Heat shock protein 80, GRP78, BIP	a
hsp90	Heat shock protein 90	a
hsp100	Heat shock protein 100, Endoplasmin	a
IFa	Intermediary filament associated	a
k8	Cytokeratin 8	b and a
LamB	Lamin B	a
Lip1	Lipocortin I	a
Lip2	Lipocortin II	a
Lip5	Lipocortin V	a
Mit1	Mitcon 1/β – F1 ATPase	a
Mit2	Mitcon 2	a
Mit3	Mitcon 3	a
MRP	Mucine Related Polypeptides	–
pcna	Ploliferating cell nuclear antigen	c and a
PLC	Phospholipase C (I)	a
RO	RO/SS-A antigen	a
SA	Serum Albumin	b and a
aT	*alpha*-Tubulin	a
bT	*betha*-Tubulin	a
tm1	Non-muscle tropomyosin isoform 1	b and a
tm2	Non-muscle tropomyosin isoform 2	b and a
tm3	Non-muscle tropomyosin isoform 3	b and a
tm4	Non-muscle tropomyosin isoform 4	b and a
tm5	Non-muscle tropomyosin isoform 5	b and a
TPI	Triose phosphate isomerase	a
V	Vimentin	b and a
Vid1	Vimentin derived protein	b and a
Vid2	Vimentin derived protein	b and a
Vid3	Vimentin derived protein	b and a
Vid4	Vimentin derived protein	b and a
Vin	Vinculin	a

a. homologous position with respect to other mammalian systems
b. purified protein(s)
c. immunoblotting

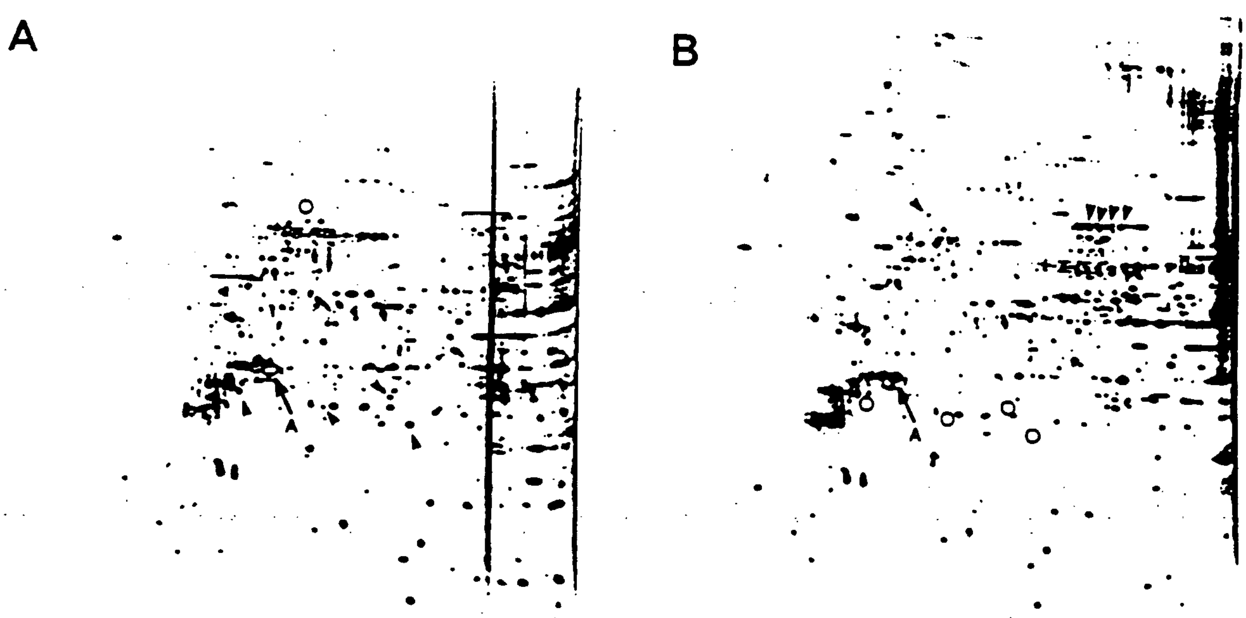


Figure 4. 2-DE analysis of a case of breast carcinoma (BC). Comparison of 2-DE quality and some differences in detected spots (arrow heads indicate increased intensity and circles or bracket indicate decreased intensity of the same spots) between (A) enzymatically and (B) nonenzymatically (scraped) tissue preparation.

3.2.2 Comparison of different methods for preparing cells from fresh tumor tissue

Samples were prepared from breast and lung carcinomas using either an enzymatic treatment with collagenase/elastase or using nonenzymatic preparations (Fig. 4). A number of differences in the protein profiles were observed in the resulting 2-DE gels, some of which are indicated in Figs. 4a and b. These differences include both increases and decreases in spot intensity. These differences may result from degradation of high molecular weight polypeptides during enzymatic treatment, increased solubilization of polypeptides, or may have other causes. For many tumors, it was only possible to obtain small amounts of material since they were reserved for other examinations. In these cases, samples could be prepared for 2-DE using either needle aspiration or scraping. Figure 5a shows a 2-DE gel prepared from squamous lung carcinoma (LS) cells collected by needle aspiration and Fig. 5b shows a gel prepared from the same tumor by scraping. In this case, a number of differences were recorded between the two procedures, some of which are arrowed in Fig. 5. Samples obtained from other tumors (breast and lung) generally showed fewer differences between these two methods of cell sampling (not shown). These data show that different nonenzymatic extraction procedures may yield different polypeptide patterns. However, the number of spots with a large

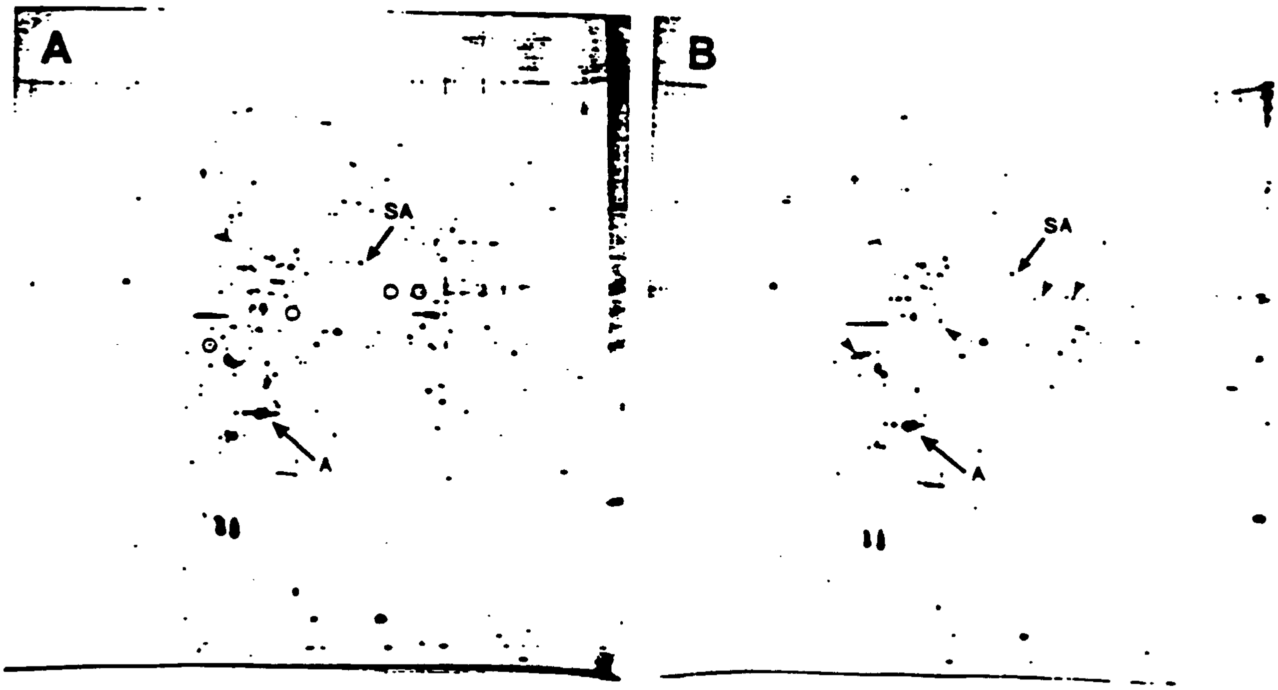


Figure 5. 2-DE analysis of a case of lung cancer (LS). Comparison of 2-DE gel quality and detected spots (arrow heads and circles) between (A) aspirated (needle aspiration) and (B) scraped preparations from fresh tissue.

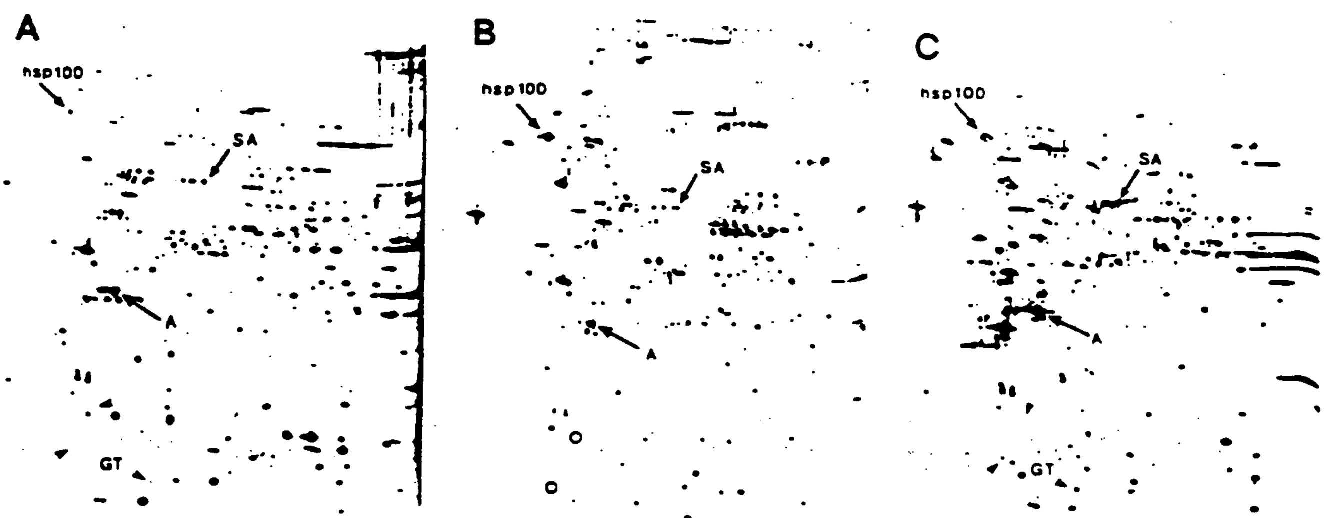


Figure 6. 2-DE analysis of three other types of tumors. (A) hypernephroma, (B) an adenoma of the thyroid and (C) corpus cancer, using the nonenzymatic preparation technique. Arrowheads and circles indicate some cytosolic polypeptides.

difference in intensity were lower than when a nonenzymatic preparation was compared with an enzymatic preparation.

2-DE maps of satisfactory quality were prepared by a third procedure. Cells were released from small pieces of tumor by squeezing (see Section 2). Some examples of this are shown in Fig. 6 where 2-DE maps derived from a case of hypernephroma, KH (Fig. 6a), a case of thyroid tumor, TA (Fig. 6b) and a case of corpus cancer, CP (Fig. 6c) can be seen. We conclude that nonenzymatic techniques are useful for 2-DE analysis of a number of different tumors. The quality of the resulting gels is comparable to that obtained using cultured cells (compare the gels in Fig. 2 with those in Fig. 4, 6 and 7). Which of these methods will be optimal will, in our experience, depend on the tumor material. For example, very small tumors are preferably extracted by squeezing; on the other hand, breast cancers (which are often fibrous) yield satisfactory samples using scraping.

3.2.3 Purification of cells on percoll gradients

We considered the possible advantage of separating viable cells from dead cells, erythrocytes, and debris using discontinuous Percoll gradients. Cells collected

Figure 7. 2-DE analysis of polypeptides from viable (b and d) and nonviable (a and c) cells of an adenocarcinoma of the lung (LB), separated using discontinuous Percoll density gradient. Nonenzymatic preparation technique (a and b) and enzymatic preparation technique (c and d) are compared.

from the interphase showed a viability of more than 90% as judged by trypan blue exclusion test. However, it as found that the yield of viable cells decreased dramatically if the tissue resection was not immediately processed. To study the effect of lysis of cells during the preparation procedure, 2-DE maps were prepared from nonenzymatically extracted cells of case LB collected from the top fraction (nonviable, Fig. 7a) and interphase fraction (viable, Fig. 7b). These 2-DE maps were compared with corresponding fractions (nonviable, Fig. 7c, and viable, Fig. 7d) of enzymatically extracted cells. One clear disadvantage of the enzymatic technique was that when loss of cell viability occurred during preparation, a dramatic loss of high molecular weight polypeptides was observed (Fig. 7c). This was probably due to degradation of intracellular proteins. However, nonenzymatic preparations showed fewer differences between viable and nonviable cells: The most pronounced alteration was a decrease of a group of mucine related proteins (Fig. 7b). We conclude, therefore, that discontinuous Percoll gradient is necessary after enzymatic extraction of cells, but can be omitted from the nonenzymatical tumor sample preparation procedure.

We used the MDA-231 cell line to study the effects of cell lysis and leakage of cytosolic polypeptides during sample preparation. Remarkably, after 30, 50, 80 and 140 min of incubation in PBS/PIH at 0°C, no significant changes were observed in the 2-DE pattern (not shown). Although loss of cell viability may not result in protein degradation when cells are incubated in the presence of protease inhibitors, loss of cytosolic proteins would be expected during pelleting of cells. We monitored the loss of lactate dehydrogenase (LDH) activity into the supernatant during incubation in PBS of MDA-231 and MCF-7 breast cancer cells at 20°C. In both cases, loss of viability was paralleled by release of LDH from the cells (Fig. 8). After 5 h, 70% of the MCF-7 cells, but only 30% of the MDA-231 cells were dead (not shown).

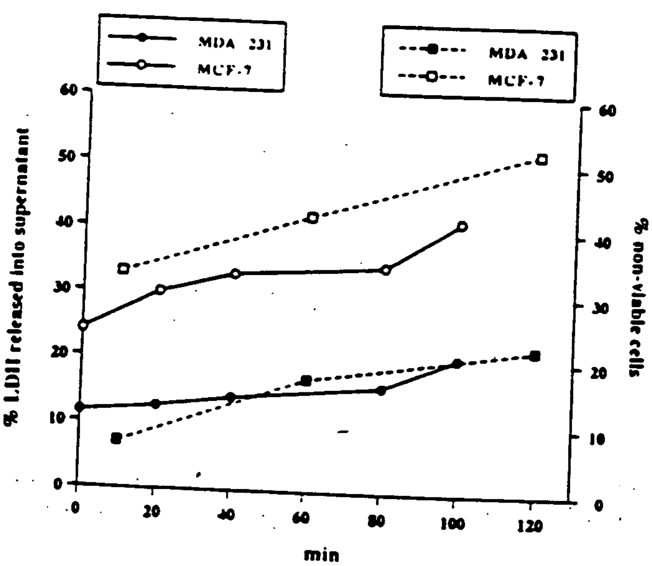


Figure 8. The relative release (fraction in supernatant of total) of lactate dehydrogenase activity (LDH) and cella viability *versus* incubation time of the mammary carcinoma cell lines MDA-231 and MCF-7 during incubation in PBS at 20°C.

These data indicate the impact of a rapid preparation procedure, at low temperature, of fresh tumor samples. Experiments have also been performed using only 1.07 g/mL Percoll (Fig. 6c and Fig. 1, left test tube) in order to remove erythrocytes. One clear advantage with this procedure, which today is routinely utilized, is a higher yield of viable cells, probably due to decreased sample preparation time.

4 Discussion

We describe procedures for sample preparation from solid tumors for 2-DE. 2-DE maps could be derived from solid tumors which were similar in quality to those obtained from cultured cells. Compared to methods using frozen material, the resolving power of the 2-DE technique is increased, allowing examination of a large number of polypeptides from tumors of different malignancies. Other investigators [12,22] have used samples from frozen tumors to derive 2-DE maps. We have previously described disadvantages encountered using frozen tumor samples including variations in contaminating proteins between different samples [3]. The methods described here are based on the preparation of cells from tumors without enzymatic digestion. The enzymatic step could be avoided since malignant cells usually grow as solid masses which are not strongly attached to the matrix. Furthermore, we found that omitting the enzymatic digestion alleviated the necessity of purifying viable tumor cells on Percoll gradients. This was in sharp contrast to enzymatically treated samples, where loss of viability leads to loss of high molecular weight proteins (Fig. 7c).

At least in the case of lung cancer, viable and nonviable cells showed small differences in respect to 2-DE maps. Presumably, protease inhibitors penetrate cells and inhibit proteolysis. In model experiments, we observed leakage of cytosolic protein (LDH) from the cells in parallel to loss of viability. Apparently, however, only a limited decrease of the level of low molecular weight cytosolic polypeptides was detected using silver staining combined with visual inspection. We have found that although some tumors are well suited for the preparation procedure described, others are not. In general, good results were obtained using tumors of the lung, breast, corpus and lymphomas. In contrast, cells from thyroid adenomas and hypernephroma showed poor viability. We were in these cases unable to separate nonviable cells from viable cells, and we can therefore not evaluate the consequence of the loss of viability on 2-DE patterns, apart from a loss of some low molecular weight cytosolic polypeptides.

Highly differentiated tumors may show lower viability as compared with poorly differentiated tumors (Dr. Farkas Vanky, personal communication). A number of samples from thyroid tumors were prepared for 2-DE but most cases showed poor viability. We believe that special care is needed during preparation of generally highly differentiated tumor groups. The difference between loss of viability/leakage of LDH of the more differentiated MCF-7 cells and the less differentiated MDA-231 cells is in line

with these observations (Fig. 8). A number of potential and interesting markers, like tropomyosin isoforms, cytokeratins and heat shock proteins, appear to be insensitive to loss of viability during the preparation procedure. We have to date made numerous observations of alterations in the expression of these polypeptides in breast cancers and lung cancers.

Another problem that may occur, irrespective of sample preparation techniques used, is admixture of lymphocytes. These cases are easily detectable in smears and it may therefore be possible to select lymphocyte specific spots as "internal markers" for the 2-D PAGE analysis. Studies using this approach are in progress. Many of the polypeptides identified are structural (Table 1). Since the expression of many of these polypeptides are known to vary between normal and malignant cells, the possibility to determine their expression simultaneously is appealing. In the specific case of breast cancer, alterations in the expression of intermediate filament proteins (cytokeratins) are known to occur during tumor progression [23]. Other proteins known to be differentially expressed between normal cells and transformed cells are tropomyosins, numatrin/B23, heat shock proteins and PCNA. To this end, we have observed alterations in the expression of cytokeratin 8, hsp 90, and non-muscle tropomyosin isoform 2 during malignant progression. (Okuzawa *et al.*, in preparation and Franzén *et al.*, in preparation).

The method of choice for sample preparation from tumor tissues will depend on the properties of the tumor material studied. It may be important to use only one method when comparing cases within one group, as differences were observed between methods. The advantages of the nonenzymatic techniques are (i) that it minimizes contamination with connective tissue, (ii) that problems with contamination of serum proteins are avoided, and (iii) that separation of viable and dead cells is not necessary. Hereby the revolving power of 2-D PAGE is maximized for the analysis of human tumors and studies on inter-tumor variations in gene expression are facilitated. In addition, the polypeptide patterns obtained may be more representative for the *in vivo* tumor cell since the use of enzymes and incubations have been minimized.

We would like to thank Dr. J. I. Garrels, Dr. S. Pattersson, Dr. S. M. Hanash and Dr. J. E. Celis for making sample and 2-DE map exchanges possible. This study was supported by grants from the Swedish Cancer Society and the Cancer Society in Stockholm.

Received March 5, 1993

5 References

[1] Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honore, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B. and Ratz, G., *Electrophoresis* 1990, *11*, 1072–1113.
[2] Garrels, J. I., Franza, B. R., Chang, C., Latter, G., *Electrophoresis* 1990, *11*, 1114–1130.
[3] Franzén, B., Iwabuchi, H., Kato, H., Lindholm, J. and Auer G., *Electrophoresis 1991*, *12*, 509–515.
[4] Sherwood, E. R., Berg, L. A., Mitchell, N. J., McNeal, J. E., Kozlowski, J. M. and Lee, C., *J. Urology* 1990, *143*, 167–171.
[5] Endler, A. T., Young, D. S., Wold, L. E., Lieber, M. M. and Curie, R. M., *J. Clin. Chem. Clin. Biochem. 1986*, *24*, 981–992.
[6] Forchhammer, J. and Macdonald-Bravo, H., in: Celis, J. E. and Bravo, R., (eds.), *Gene Expression in Normal and Transformed Cells*, Plenum, New York 1983, pp. 291–314.
[7] Linder, S., Brzeski, H. and Ringertz, N. R. *Exp. Cell. Res.* 1979, *120*, 1–14.
[8] Celis, J. E. and Bravo, R. (Eds.), *Two-dimensional Gel Electrophoresis of Proteins*, Academic Press, New York 1984, pp. 3–36.
[9] Garrels, J. I., *J. Biol. Chem.* 1979, *254*, 7961–7977.
[10] Anderson, N. L., *Two-Dimensional Electrophoresis. Operation of the ISO-DALT System*, Large Scale Biology Press, Washington, DC 1988, 162.
[11] Bradford, M., *Anal. Biochem.* 1976, *72*, 248.
[12] Tracy, R. P., Wold, L. E., Currie, L. M. and Young, D. S., *Clin. Chem.* 1982, *28*, 890–899.
[13] Merril, C. R., Goldman, D., Sedman, S. A. and Elbert, H. M., *Science* 1981, *211*, 1437–1438.
[14] Morrissey, J. H., *Anal Biochem.* 1981, *117*, 307–310.
[15] Gard, D. L., Bell, P. B., Lazarides, E., *Proc. Natl. Acad. Sci. USA*, 1979, *76*, 3894–3898.
[16] Matsumura, F., Lin, J.-C., Yamashiko-Matsumura, S., Thomas, G. P. and Topp, W. C., *J. Biol. Chem.*, 1983, *258*, 13954–13960.
[17] Paulin, D., Forest, N. and Perreau, J., *J. Mol. Biol.* 1980, *144*, 95–101
[18] Blobel, G. A., Moll, R., Franke, W. W., Kayser, K. W. and Gould, V. E., *Am. J. Pathol.* 1985, *121*, 235–247.
[19] Ochs, D. C., McConkey, H. E. and Guard, N. L., *Exp. Cell. Res.* 1981, *135*, 355–362.
[20] Bhattacharya, B., Gaddamanuga, L.P., Valverius, E. M., Salomon, D. S. and H. L. Cooper, *Cancer Res.* 1990, *50*, 2105–2112.
[21] Sommers, C. L., Walker-Jones, D., Heckford, S. E., Worland, P., Valverius, A., Clark, R., McCornick, F., Stampfer, M., Abularch, S. and Gelmann, E. P., *Cancer Res.* 1989, *49*, 4258–4263.
[22] Trask, D. K., Band, V., Zajchwski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, *87*, 2319–2323.
[23] Trask, D. K., Bond, V., Zajchanski, D. A., Yaswen, P., Suh, T. and Sager, R., *Proc. Natl. Acad. Sci. USA* 1990, *87*, 2319–2323.

Docket No.: PF-0232-1 DIV
USSN [redacted] 782,390
Ref. [redacted]

Bengt Bjellqvist*
Bodil Basse
Eydfinnur Olsen
Julio E. Celis

Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, Aarhus

Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions

A highly reproducible, commercial and nonlinear, wide-range immobilized pH gradient (IPG) was used to generate two-dimensional (2-D) gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human epidermal keratinocytes. Forty one proteins, common to most human cell types and recorded in the human keratinocyte 2-D gel protein database were identified in the 2-D gel maps and their isoelectric points (p*I*) were determined using narrow-range IPGs. The latter established a pH scale that allowed comparisons between 2-D gel maps generated either with other IPGs in the first dimension or with different human protein samples. Of the 41 proteins identified, a subset of 18 was defined as suitable to evaluate the correlation between calculated and experimental p*I* values for polypeptides with known composition. The variance calculated for the discrepancies between calculated and experimental p*I* values for these proteins was 0.001 pH units. Comparison of the values by the *t*-test for dependent samples (paired test) gave a *p*-level of 0.49, indicating that there is no significant difference between the calculated and experimental p*I* values. The precision of the calculated values depended on the buffer capacity of the proteins, and on average, it improved with increased buffer capacity. As shown here, the widely available information on protein sequences cannot, *a priori*, be assumed to be sufficient for calculating p*I* values because post-translational modifications, in particular *N*-terminal blockage, pose a major problem. Of the 36 proteins analyzed in this study, 18–20 were found to be *N*-terminally blocked and of these only 6 were indicated as such in databases. The probability of *N*-terminal blockage depended on the nature of the *N*-terminal group. Twenty six of the proteins had either M, S or A as *N*-terminal amino acids and of these 17–19 were blocked. Only 1 in 10 proteins containing other *N*-terminal groups were blocked.

1 Introduction

As compared with carrier ampholyte isoelectric focusing (CA-IEF), the application of immobilized pH gradients (IPGs) in the first dimension in 2-D gel electrophoresis offers improved reproducibility [1] because the nature of the pH gradient makes the resulting focusing positions insensitive to the focusing time [2] and to the type of sample applied [3]. The recently introduced ready-made IPG strips [4] seem to be an ideal substitute for the carrier ampholyte gradients, which until now have been the most commonly used first dimensions in 2-D gel electrophoresis. The availability of standardized first dimensions opens the possibility of comparing 2-D gel maps of various cell types generated in different laboratories, provided that the focusing positions of a number of easily recognizable polypeptide spots common to the cell types in question are known. Even though this approach is limited to experiments performed with the same standardized IPG, the flexibility provided by IPGs allows the pH gradient to be adjusted to the requirements of a particular experiment.

Exchange and communication of 2-D gel protein data requires a pH scale that is independent of the particular IPG used and by which the results can be described. The introduction of carbamylation trains and the relation of focusing positions to the spots in these trains represented a step forward towards solving the reproducibility problem experienced with carrier ampholyte focusing [5]. Problems associated with the use of carbamylation trains were mainly due to lack of temperature control and to the use of nonequilibrium focusing conditions. Accordingly, the pattern variation involved not only the resulting pH gradients, but also the relative spot positions as related to each other and to spots in the carbamylation trains. Even though the question of reproducibility has, to a large extent, been solved, the carbamylation trains are still not ideal as markers because the spots in the trains do not represent defined entities but rather a large number of differently carbamylated peptides having close p*I* values. As a result, the spots are large and poorly defined as compared to the ordinary polypeptide spots in 2-D gel maps.

Correspondence: Professor J. E. Celis, Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Aarhus University, DK-8000 Aarhus C, Denmark

Abbreviations: CA-IEF, carrier ampholyte-isoelectric focusing; **SSP,** sample spot number

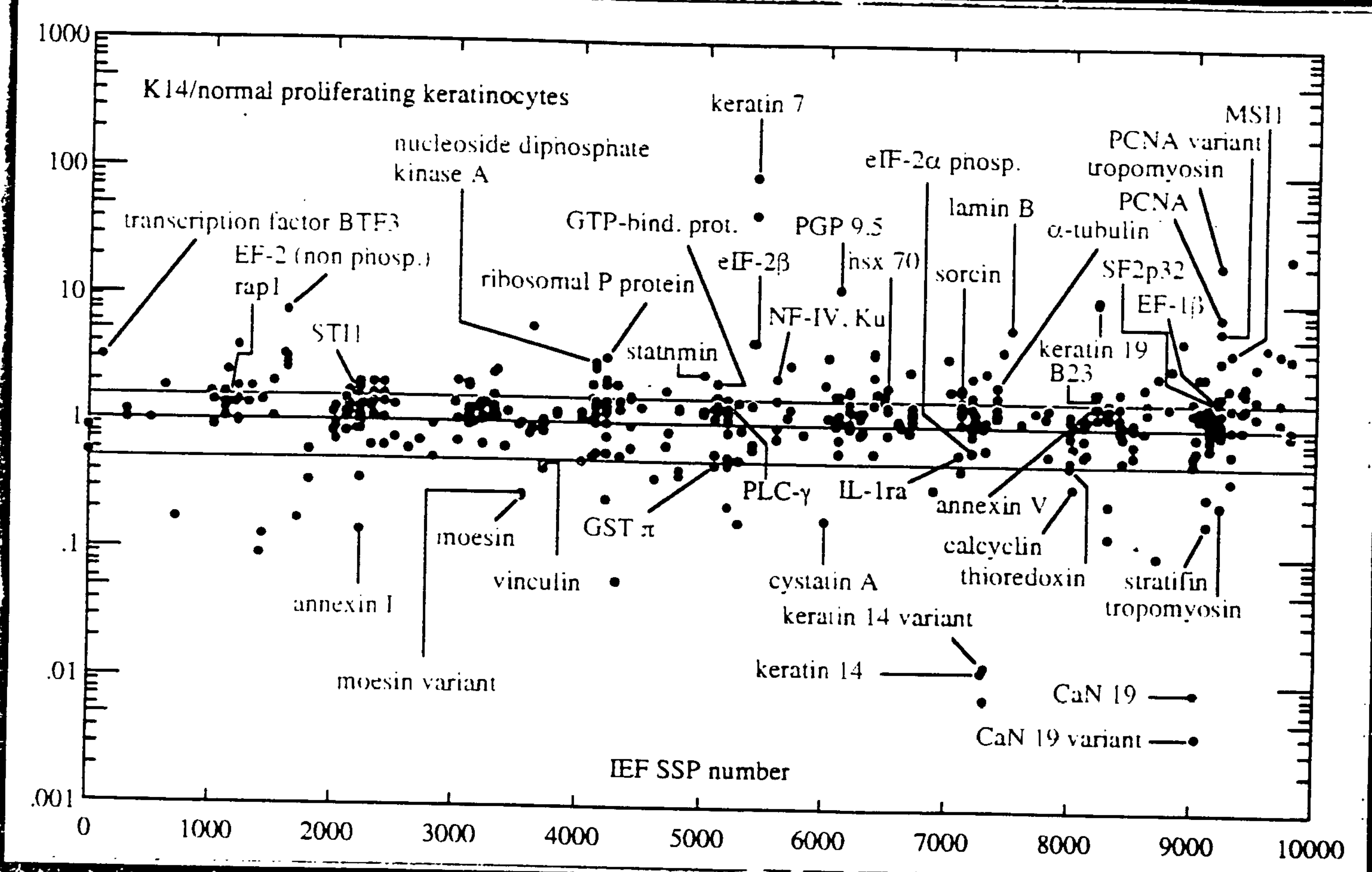
* Present addess: Pharmacia Biotech AB, S-751 82 Uppsala, Sweden

 0173-0835/94/0304-0529 $5.00+.25/0

ELECTROPHORESIS

An International Journal

3-4'94



PAPER SYMPOSIUM

ELECTROPHORESIS IN CANCER RESEARCH

Guest Editor: Julio E. Celis

VCH

ISSN 0173-0835 · ELCTDN 15 (3-4) 307–556 (1994) · Vol. 15 · No. 3-4 · March/April 1994

Electrophoresis 1993, 14, 526–536

Neidhardt *et al.* [6] defined the pH gradient in 2-D gel experiments by p*I* markers whose p*I* values were calculated from the amino acid composition. Focusing positions of other polypeptides could be predicted from their composition but the p*K* values needed for the p*I* calculations were unknown. Various groups employing this approach do not use the same pK values [6, 7] and therefore, the p*I* values derived in this way cannot be expected to describe the variation of the hydrogen ion activity. In spite of this fact, it is still possible to make approximate predictions of focusing positions because the p*K* values used to define the pH gradient are also used to calculate p*I* values and to predict the focusing positions. Errors in p*K* assignments are therefore compensated. A pH scale which corretly reflects the variation in hydrogen ion activity during focusing should improve the precision of the predictions, but this has never been implemented with CA-IEF focusing as a first dimension in 2-D gel electrophoresis. The main reason for this are the problems associated with pH measurements in focused gels containing high concentrations of urea.

IPGs can be described from the concentration variation of the immobilized groups, provided that the p*K* values of these groups are known for the conditions prevailing during focusing. To avoid measurements on gels, Gianazza *et al.* [8] suggested the use of p*K* values derived by addition of determined p*K* shifts. Recently, direct determinations of p*K* differences between immobilized groups in IPGs were made by determining p*I*-p*K* values in overlapping narrow-range IPGs [9, 10] and the results verified the applicability of the Gianazza approach. A description of the focusing results in a pH scale, which correctly describes the variation of the hydrogen ion activity for the focusing conditions used, not only allows the comparison of 2-D gel maps generated with different IPGs, but also opens the possibility for correlating the focusing position of a polypeptide with its composition [9]. Experiments by Bjellqvist *et al.* [9, 10] have implied that pH scales showing good correlation between calculated and experimental p*I* values can be derived for any of the conditions commonly used for focusing in connection with 2-D gel electrophoresis. These pH scales are then defined through the p*K* values of the immobilized groups in the IPG containing gel. To be useful for interlaboratory comparisons, however, the pH scale has to be defined through p*I* values of easily recognizable spots present in the 2-D gel map. So far, p*I* determinations in a useful pH scale, combined with determinations of pK values needed for p*I* calculations, have only been made for the pH range 4.5–6.5 at 10°C [9]. CA-IEF focusing as described by O'Farrell [11] does not control the temperature of the first dimension, which can be expected to be slightly above room temperature. With IPGs, the temperature commonly used is about 20°C [4, 12] or 25°C [13] and this is a critical parameter that needs to be controlled [14].

The present work was designed to compare 2-D gel maps of different cell types in a laboratory applying both CA-IEF and IPG focusing at a common temperature. To this end we have generated 2-D gel maps of proteins from noncultured, unfractionated normal human epidermal keratinocytes with IPG in the first dimension and a focusing temperature of 25°C. We have used commercial nonlinear, wide-range IPG strips which give 2-D gel maps that are closely similar to the ones resulting with the CA-IEF technique used to establish the human keratinocyte database [15]. As an initial step towards interlaboratory comparisons of results obtained with the nonlinear gradient as a first dimension we report here on the focusing positions of 41 known proteins that are common to most human cell types. The pH range covered corresponds to the range in classical CA-IEF 2-D gel electrophoresis and in order to use these proteins as internal standards for comparing 2-D gel maps generated with other IPGs we determined their p*I* values with narrow-range IPGs in the first dimension. We have compared the calculated *versus* experimental p*I* values and show that it is necessary to have further information (absence or presence and nature of posttranslational modifications), in addition to amino acid composition to be able to calculate p*I* values that correspond to the actual experimental values. The p*K* values used for the calculations are provided and the usefulness of p*I* prediction in relation to database information is discussed. Furthermore, we comment on the possibility of using experimentally determined p*I* values to verify the available database information on polypeptide composition.

2 Materials and methods

2.1 Apparatus and chemicals

Equipment for isoelectric focusing and horizontal SDS electrophoresis (Multiphor® II electrophoresis chamber, Immobiline® strip tray, Multidrive XL programmable power supply, Macrodrive power supply and Multitemp® II) was from Pharmacia LKB Biotechnology AB (Uppsala, Sweden). Vertical second-dimensional gels were run in the home-made equipment described in [15]. The IPG strips with the wide-range nonlinear pH gradient were either Immobiline DryStrip® pH 3–10 NL, 180 mm or alternatively 160 mm long IPG strips with a corresponding pH gradient. In both cases the IPG strips were delivered by Pharmacia LKB. Immobiline, Pharmalyte, Ampholine, GelBond as well as PAG film and the ready-made horizontal SDS gels (ExcelGel® XL SDS 12–14) were also from Pharmacia LKB. Purified proteins and peptides were from Sigma (St. Louis, MO).

2.2 Sample preparation

Preparation and labeling of unfractionated keratinocytes as well as fibroblasts have been described in [16]. Cells were lysed in a solution containing 9.8 M urea, 2% w/v NP-40, 100 mM DTT and 2% v/v Ampholine pH 7–9.

2.3 2-D gel electrophoresis

First-dimensional focusing was performed according to Görg *et al.* [2] with some minor modifications, as described in [9]. Rehydration of the IPG strips was made in a solution containing 9.8 M urea, 2% w/v CHAPS, 10 mM DTT and 2% v/v carrier ampholyte mixture. The carrier ampholyte mixture consisted of 2 parts Pharmalyte

4–6.5, 1 part Ampholine pH 6–8 and 1 part Pharmalyte pH 8–10.5. Usually, cathodic sample application was used and the samples were diluted 2–20 times in a solution containing 9.8 M urea, 4% w/v CHAPS, 1% w/v DTT and 35 mM Tris base. For acidic application, the Tris-base was substituted with 100 mM acetic acid. The degree of dilution and sample volume (20–100 µL) depended on the particular sample and the IPG, and whether visualization of the proteins was to be done by Coomassie Brilliant Blue or silver staining. With the wide-range non-linear IPG, 10–30 µg of total protein was loaded for silver staining and 100–200 µg for Coomassie staining. Focusing was done overnight with Vh products in the range of 45–60 kVh with 160 mm long strips and 50–70 kVh with 180 mm long strips. Solubilization of polypeptides and blocking of -SH groups prior to the second-dimensional run, as well as loading on the second-dimensional gel was done as described in [9]. The stacking gel was omitted and 5–10 mm were left at the top of the second-dimensional gel for applying the IPG strip. The space was filled with electrode buffer containing 0.5% w/v agarose. Casting, running, staining and autoradiography were carried out as described in [15].

2.4 Experimental determination of p*I* values

The determination of the p*K* differences between Immobilines p*K* 4.6, p*K* 6.2 and p*K* 7.0 necessary for the calibration of the pH scale at 25°C in 9.8 M urea was done as described in [9] with the same narrow-range IPGs. The pH scale was defined by setting the p*K* value of Immobiline p*K* 4.6 equal to 4.61 [9] and the determined p*K* differences gave the p*K* values of Immobilines p*K* 6.2 and p*K* 7.0, equal to 5.73 and 6.54, respectively. The p*K* differences found are in good agreement with values derived from [17] and [8] by extrapolation to 9.8 M urea concentration. As in [9], additional narrow-range recipes have been used for determining p*I* values. With narrow-range IPGs extending to pH values higher than the p*K* value of Immobiline p*K* 7.0, anodic sample application was used with acetic acid added to the sample solution. Otherwise, cathodic sample application was used with the same sample buffer as for wide-range IPGs.

2.5 Protein compositions used for p*I* calculations

With the exception of vimentin, protein compositions are from the Swiss-Prot database [18]. For vimentin, we used the data from [19], where the amino acid at position 41 is a D instead of a S. Information in the Swiss-Prot database on phosphorylation has been disregarded because it was known from earlier studies (J. E. Celis, unpublished results) that the spots in question corresponded to the unphosphorylated forms of the peptides.

2.6 Calculation of p*I* values

For the p*I* calculations it was assumed that the same p*K* value could be used for an amino acid residue in all polypeptides and in all positions in the peptide except for *N*- or *C*-terminally placed amino acids. For the p*K* values of the *N*-terminal amino groups the effect of the different substituents on the α-carbon were taken into account. The calculations of p*I* values were made with the aid of the IPG-maker program [20].

2.7 p*K* values used for p*I* calculations

For the carboxyl terminal group and internal glutamyl and aspartyl residues the same p*K* values were used as in [9]. For *C*-terminal glutamyl and aspartyl residues, separate p*K* values were derived with the aid of the Taft equations [9, 21]. The p*K* values of histidyl groups were calculated from the p*I* values of human carbonic anhydrase I as in [9]. For *N*-terminal glycine a p*K* value of 7.50 was used. The p*K* shift caused by a substituent on the α-carbon was assumed to be identical with the p*K* shift the substituent caused for the amino group in the amino acid, *i.e.* 2.28 pH units were subtracted from the p*K* values for the amino groups in the amino acids given in [22, 23]. The approximate p*K* value of 9 for the cystenyl group was taken from [24]. For tyrosyl and arginyl groups we used the p*K* values for the amino acids [22, 23]. For lysyl groups the effect of high urea concentration on amino groups was taken into account and 0.5 pH units were subtracted from the amino acid p*K* value. These last three p*K* values are far from the pH range under study and the results found would have been the same if lysyl and arginyl groups were assumed to be fully ionized while the ionization of tyrosyl groups were neglected. A complete list of the p*K* values used is given in Table 1.

Table 1. p*K* Values used for the ionizable groups in peptides 9.8 M urea, 25°C

Ionizable group	p*K*
C-terminal	3.55*
N-terminal	
Ala	7.59
Met	7.00
Ser	6.93
Pro	8.36
Thr	6.82
Val	7.44
Glu	7.70
Internal	
Asp	4.05
Glu	4.45
His	5.98
Cys	9
Tyr	10
Lys	10
Arg	12
C-terminal side chain groups	
Asp	4.55
Glu	4.75

2.8 Statistical analysis

Statistical comparisons of the experimental and calculated p*I* values were done on an Apple Macintosh IIsi using the statistical package Statistica/Mac, release 3.0b (from StatSoft Inc., Tulsa, Oklahoma). Calculated and experimental p*I* values were compared by the *t*-test for

correlated samples (paired t-test). The normality of pI differences was estimated graphically by probability plots. The variances of the data presented here and the similar data on plasma and liver proteins in [9] were compared by the F-test.

3 Results and discussion

3.1 Identification of polypeptides and pI determinations

The 2-D gel maps of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes, focused with the nonlinear, wide-range IPG and CA-IEF pH gradients in the first dimension, are shown in Figs. 1 and 2, respectively. The IPG extends to higher pH values but otherwise the two patterns are very similar and most of the spots in the IPG pattern can be directly related to the corresponding spots in the CA-IEF gel. To obtain comparable patterns it was important to keep the focusing temperature as similar as possible. Compared to other studies [1–4, 9, 10, 12–14], we increased the urea concentration in the focusing gel to 9.8 M because keratins streaked badly in the focusing dimension when 8 M urea was used, presumably due to

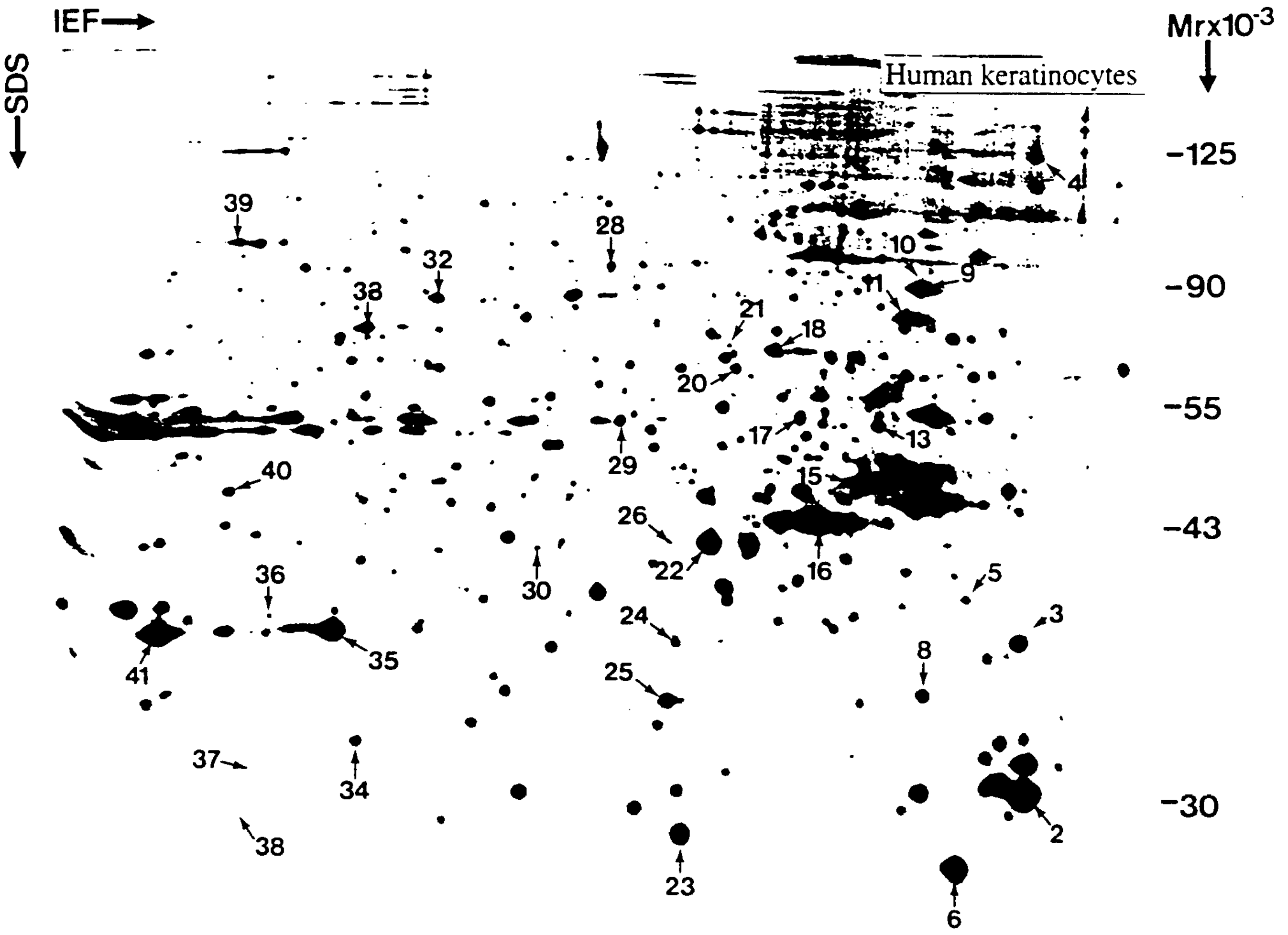


Figure 1. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

aggregates of acidic and basic keratins. An increase in urea concentration to 9 M or more eliminated these streaks; apart from this effect, no other major changes in the focusing positions were observed. In Fig. 1 we have indicated the positions of 41 known proteins from the human keratinocyte 2-D gel database that are most likely common to most human cell types. The choice was made because these proteins are easy to identify with certainty. With the exception of stratifin (spot 2), involucrin (spot 4) and keratin 14 (spot 15), which are all epithelial markers, these proteins are also present in human fibroblasts (Fig. 3) and lymphocytes (results not shown), and therefore can be used as landmarks for comparing 2-D gel maps derived from different cell types. In Table 2 the 41 proteins are listed together with their sample spot numbers (SSP) in the human keratinocyte protein database and p*I* values determined in 2-D gel maps generated with narrow-range IPGs in the first dimension.

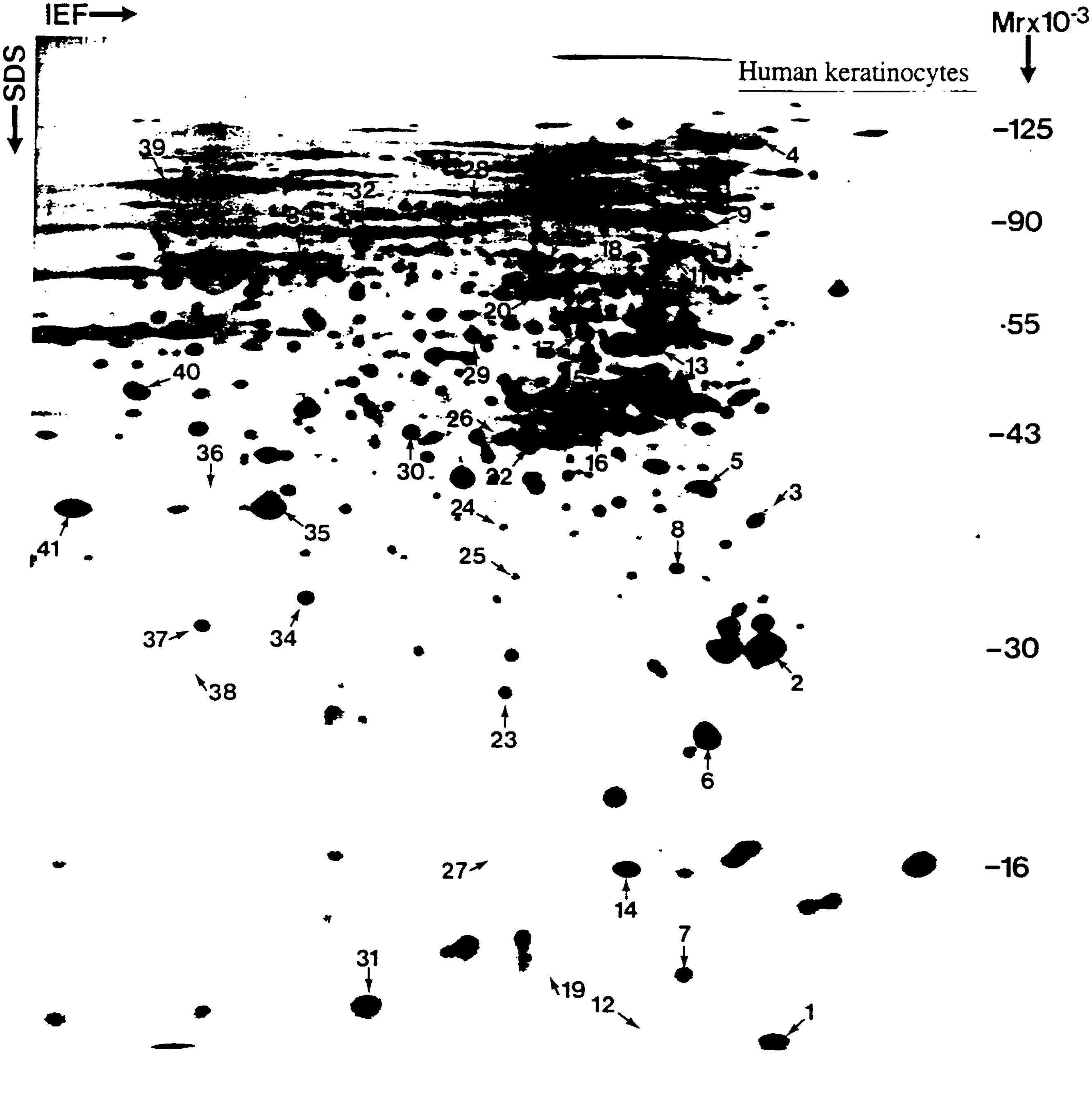


Figure 2. 2-D gel protein map of [^{35}S]methionine-labeled proteins from noncultured, unfractionated normal human keratinocytes focused with CA-IEF in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

Table 2. Proteins from the human keratinocyte database localized in 2-D gels run with IPGs as first dimension

Number in Figs. 1–3	Protein name	IEF SSP number[a]	Experimental pI value	Calculated pI value	Discrepancy (pH units)	Calculated net charge at experimental pI value	Buffer capacity charge units pro pH unit	N-terminal	Recalculated for suspected blockage		N-terminal	Swiss-Prot accession number
									pI value	Discrepancy pH units	Net charge	
1	CaN 19	9027	4.46	–	–	–	–	–				–
2	Stratifin, bovine 14-3-3 related protein	9109	4.58	–	–	–	–	–				–
3	Proliferating nuclear antigen (PCNA)/cyclin	9226	4.58	4.57	–0.01	–0.1	20.8	M				P12004
4	Involucrin	9703	4.63	4.63	0.00	–0.3	70.1	M				P07476
5	Nucleolar protein B23	8207	4.75	4.64	–0.11	–3.2	30.4	M				P06748
6	Translationally controlled tumor protein	8114	4.79	4.84	0.05	0.6	13.1	M[b]				P13693
7	Thioredoxin	8006	4.86	4.82	–0.04	–0.3	7.1	V[b]				P10599
8	Annexin V	8213	4.89	4.88	–0.01	–0.1	20.3	A[c]				P08758
9	Heat shock protein 90-β	8611	4.95	4.94	–0.01	–0.5	56.2	P				P07900
10	Heat shock protein 90-α	2629	4.97	4.97	0.00	0.2	53.6	P				P08238
11	Glucose regulated protein 78 (BiP)	8515	4.99	4.98	–0.01	–0.6	37.5	E				P11021
12	Calcyclin	8017	5.02	5.32	0.30	1.3	3.6	M	5.09	0.07	0.3	P06703
13	Vimentin	8417	5.05	5.06	0.01	0.2	27.1	S				P08670
14	Initiation factor 4D	8016	5.05	5.08	0.03	0.2	7.6	A[c]				P10159
15	Keratin 14	7305	5.08	5.09	0.01	0.2	21.0	T				P02533
16	β-Actin	7316	5.21	5.21	0.00	0.06	13.3	D[c]				P02570
17	Heat shock protein 60	6403	5.23	5.24	0.01	0.1	17.5	A[b]				P10809
18	Heat shock cognate 71kD	6504	5.28	5.37	0.09	1.8	18.1	M	5.32	0.04	0.8	P11142
19	Cystatin	6011	5.30	5.38	0.08	0.2	3.0	M				P01040
20	T-plastin	6412	5.34	5.41	0.07	1.3	17.7	M	5.36	0.02	0.3	P13797
21	Calelectrin	5628	5.35	5.37	0.02	0.5	23.3	A[c]				P08133
22	Plasminogen activator inhibitor-2	6314	5.38	5.46	0.08	0.9	10.7	M	5.37	–0.01	–0.07	P05120
23	Glutathione S-transferase π	5101	5.43	5.44	0.01	0.08	3.9	P				P09211
24	Annexin VIII	5213	5.45	5.56	0.11	1.0	8.7	M	5.46	0.01	0.05	P13928
25	Annexin III	5204	5.46	5.63	0.17	1.4	8.4	M	5.52	0.06	0.5	P12429
26	Adenosine deaminase	5305	5.47	5.63	0.16	1.8	10.8	M	5.54	0.07	0.8	P00813
27	Stathmin	5001	5.55	5.61	0.06	0.4	6.6	A[c]				P16949
28	Gelsolin, cytoplasmic	5608	5.59	5.58	–0.01	–0.1	16.5	V				P06396
29	Rat phosphoinoside specific protein homolog	5410	5.62	–	–	–	–	–				
30	Elastase inhibitor	4314	5.74	–	–	–	–	–				
31	S100, calgizarin	4006	5.75	–	–	–	–	–				
32	Cytvillin, ezrin	3504	5.99	5.95	–0.04	–0.5	13.2	P				P15311
33	Moesin	3515	6.11	6.09	–0.02	–0.2	9.8	P				P26038
34	Purine nucleoside phosphorylase	2108	6.11	6.45	0.34	1.8	4.4	M	6.28	0.17	0.9	P00491
35	Annexin I	2216	6.18	6.64*	0.46	1.6	2.5	A	6.33	0.15	0.6	P04083
36	Aldose reductase	1202	6.40	6.55	0.15	0.7	4.2	A	6.36	–0.04	–0.2	P15121
37	Phosphoglycerate mutase (B form)	1107	6.46	6.75	0.29	0.9	2.6	A	6.46	0.00	0.0	P18669
38	Triosephosphate isomerase	1111	6.53	6.51	–0.02	–0.04	2.3	A[b]				P00938
39	Elongation factor 2	1610	6.43	6.38	–0.05	–0.5	9.8	M				P13639
40	α-Enolase	1325	6.62	6.99	0.37	1.0	2.2	S	6.75	0.13	0.3	P06733
41	Annexin II	210	7.30	7.36	0.06	0.05	0.9	S[c]				P07355

a) SSP number in the keratinocyte database [15]
b) Peptides N-terminally sequenced as liver proteins [3]
c) Peptides given as N-terminally blocked in Swiss-Prot database

3.2 Comparison between the determined and calculated p*I* values for human keratinocyte proteins

Thirty six of the 41 proteins listed in Table 2 are found in the Swiss-Prot database. Contrary to the plasma and liver proteins used in [9], the p*I* calcuations on the proteins used in this study posed some problems that reflected the way in which they were characterized. The proteins used by Bjellqvist *et al.* [9] were either very abundant and well-characterized plasma proteins or they were identified by *N*-terminal sequencing and, therefore, the nature of the *N*-terminals (acetylated or non-acetylated) was in both cases known. The proteins used in this study have all been characterized by internal sequencing [7] and it is known that *N*-terminal acetylation occurs with high frequency in eukaryotes.

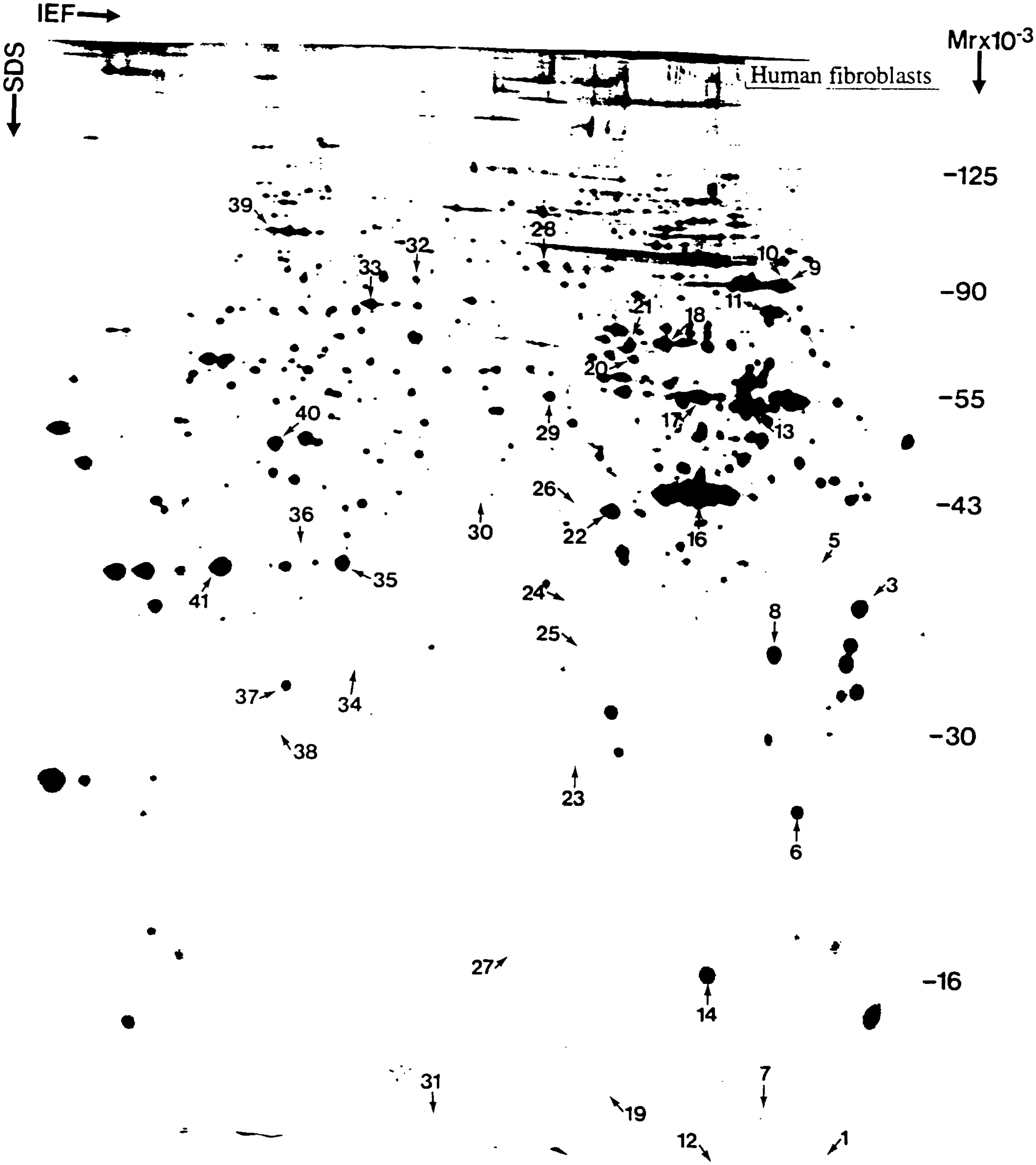


Figure 3. 2-D protein map of [^{35}S]methionine-labeled proteins from normal human fibroblasts focused with the nonlinear, wide-range IPG in the first dimension. The position of the 41 proteins analyzed in this study is indicated.

According to Brown and Robert [25], proteins with acetylated N-terminals correspond in weight to approximately 80% of the soluble protein in ascites cells. Based on results from N-terminal sequencing, at least 40% of the spots in the human liver protein 2-D gel map appear to be blocked [3]. The corresponding number, derived from 107 spots in the 2-D gel map of human T-lymphocyte proteins, falls between 60 and 65% (J. Strahler, personal communication). Information concerning N-terminal blockage is not normally available, and in the Swiss-Prot database only 6 of the 36 keratinocyte proteins are specified as N-terminally blocked. We have, within the present material, defined 18 proteins for which the N-terminals are very likely to be correctly described. Six of these proteins are listed in the Swiss-Prot database as N-terminally blocked, four represent proteins which appear in the human liver 2-D gel map and have been N-terminally sequenced as liver proteins [3] and the remaining eight have N-terminal groups other than M, S and A, *i.e.* N-terminals for which N-acetylation is uncommon [26]. In Figs. 4A, B, C and D p*I* values calculated from Swiss Prot database information are plotted against the experimentally determined p*I* values for all the keratinocyte proteins listed in Table 2 and for the 18 selected proteins, as well as for the plasma and liver proteins (data from [9] valid for 10°C)*.

The calculations show that without knowledge of the status of the N-terminal group, precise predictions of p*I* values for eukaryotic proteins cannot be achieved based on the information available in Swiss-Prot and similar databases. However, for proteins where the N-terminal status is known, we find good correlation between predicted and experimental p*I* values. When the variance of the p*I* discrepancies and the variance of calculated charges at the experimental p*I* values derived from the present data set are compared with the corresponding

* There are four plots: (A) the 36 polypeptides from normal human keratinocytes (no corrections), (B) the 36 polypeptides from Fig. 4A where p*I* values have been recalculated for 12 polypeptides with M, S and A as N-terminally assumed blocked, based on calculated charge, (C) the 18 selected polypeptides with information on the N-terminal configuration, and (D) plasma and liver proteins.

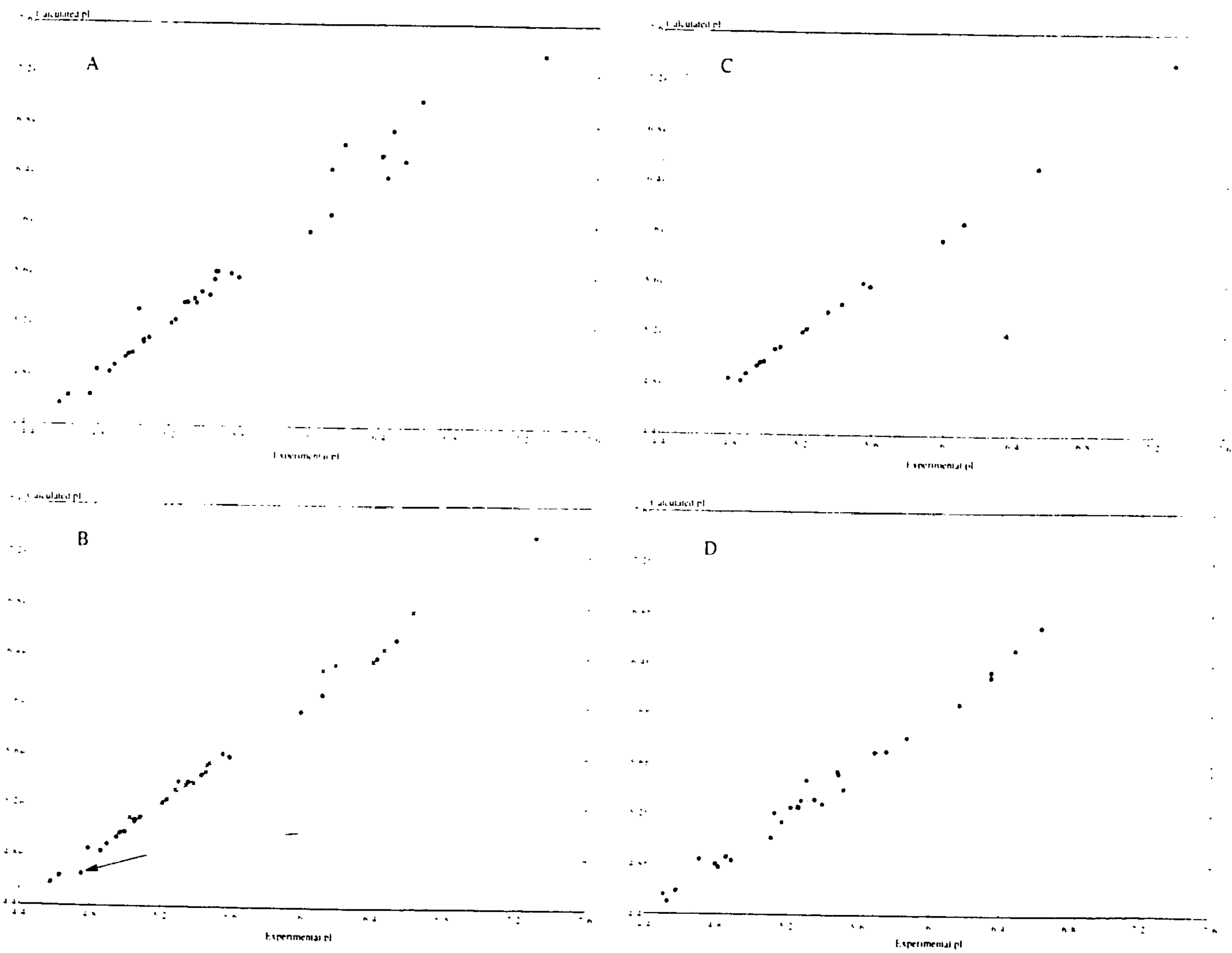


Figure 4. Calculated *vs.* experimental p*I* values. Lines are fitted using the least squares' criterion. (A) 36 polypeptides from normal human keratinocytes (no corrections). (B) 36 polypeptides from Fig. 4A (including the 18 marker polypeptides) where p*I* values have been recalculated assuming N-terminal blockage; x indicates recalculated p*I* values; nucleolar protein B23 is indicated with an arrow. (C) 18 polypeptides with information on N-terminal configuration and (D) plasma and liver proteins.

values derived from the data on plasma and liver proteins in [9] (Table 3), the present data are found to result in larger variances for the values of both p*I* discrepancies and calculated charge at the experimental p*I* value when no information on posttranslational modification is taken into consideration. Correction for possible *N*-acetylation of 12 polypeptides with M, S and A as *N*-terminal results in a smaller variance of p*I* discrepancies, although not significantly different from values derived from [9], whereas the variance of the calculated charge at the experimental p*I* value is significantly higher. For the 18 selected proteins the variance for the p*I* discrepancies is significantly smaller than for the data in [9]; however, the corresponding value for calculated charge at the experimental p*I* value does not improve to the same extent. This, we believe, reflects another difference between the two sets of proteins used for the calculations. Based on spot distributions in 2-D gel maps, the set of proteins used here has a molecular weight distribution that is more representative of the patterns observed in mammalian cells. In the study by Bjellqvist *et al.* [9] most of the high molecular weight plasma proteins had to be excluded due to their unknown content of sialic acid which made the proteins analyzed in this study heavily biased towards low molecular weight proteins. The buffer capacity of proteins normally increases with the protein's molecular weight, and the average buffer capacity of the presently selected proteins with assumed known *N*-terminals is 18 charge units/pH unit, while the corresponding value for the proteins used in [9] is only 9 charge units/pH unit. High buffer capacity can be expected to improve the agreement between calculated and experimental p*I* values. Inspection of the data presented in Table 2 for the polypeptides with assumed known *N*-terminals verifies the importance of the buffer capacity. For 8 polypeptides having buffer capacities higher than 15 charge units/pH unit, the calculations in all cases yielded p*I* discrepancies with absolute values of less than 0.02 pH units. The largest discrepancy, 0.06 pH units, was observed for annexin II and stathmin, proteins which have low buffer capacity: 0.9 and 6.6 charge units/pH unit, respectively. The probability that the focusing position of a protein with known composition will fall within a certain distance from the calculated p*I* value therefore cannot be predicted by the variance alone. The buffer capacity of the specific protein must be taken into consideration as well. As indicated by the decrease of the variance of calculated charges at the experimental p*I* value for the selected proteins, the observed improvement can not solely be due to the higher buffer capacity of the keratinocyte proteins. The two studies relate to different experimental conditions. Good agreement between experimental and calculated p*I* values implies that the proteins are defolded and a factor that may contribute to the observed improvement is a more complete defolding of proteins caused by the higher temperature and urea concentration used in this study.

The data indicated that the precision with which p*I* values can be predicted for polypeptides with high buffer capacity is better than the precision with which experimental p*I* values can be determined. If the pH is defined through the p*K* values of the immobilized groups in the IPG containing gel, the precision of the experimentally calculated data will depend on the pH difference between the p*I* and the p*K* value of the immobilized group with the closest p*K*. For the present study this will give p*I* determinations with a precision varying in the range of ± 0.02–0.05 pH units [9]. The good agreement observed between the calculated and experimental p*I* values is due to the fact that errors are mainly systematic and, as discussed in [9], they will largely be cancelled out in the calculations. A pH scale defined through the presently determined p*I* values will not necessarily reflect the variation of the hydrogen ion activity during the focusing step in an optimal way, but it still allows precise predictions of focusing positions for polypeptides with known compositions, including information on posttranslational modifications. Calculated net charge at the experimentally found isoelectric point defined in this scale will serve as a tool to verify that the polypeptide

Table 3. Mean values and variances for the difference (experimental p*I*-calculated p*I*) in pH units and calculated charges at the experimental p*I* values, respectively

	Plasma and liver proteins (8 M urea, 10°C)		Keratinocyte proteins (9.8 M urea, 25°C)					
			All peptides		All peptides after correction for *N*-acetylation		Known *N*-terminal configuration (or very likely configuration)	
Number of proteins	29		36		36		18	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Experimental p*I*-calculated p*I*	−0.011	0.005	0.072	0.017	0.019	0.003	0.005	0.001
F-value (p*I* discrepancy)[a]	1		3.4		1.67		5	
P-level (p*I* discrepancy)[b]	0.5		0.0005		0.0721		0.0004	
Calculated charge at the experimental p*I* value	−0.070	0.227	0.321	0.871	0.009	0.444	−0.014	0.109
F-value (calculated charge at the experimental p*I* value)[a]	1		3.8		1.96		2.08	
P-level (calculated charge at the experimental p*I* value)[b]	0.5		0.0002		0.0338		0.0536	

a) Comparison to the data in [9]. $F = S_1^2/S_2^2$, where S_1^2 is the larger of the two variances

b) $P(F(\nu_1, \nu_2) \geq F\text{-value})$, where ν_1 and ν_2 are the degrees of freedom for s_1 and s_2, respectively

composition used in the calculation is correct and complete. Exceptions to this are proteins such as involucrin and heat shock protein 90 that have very high buffer capacities. Introduction of an extra charge unit into these proteins will only result in p*I* shifts falling in the range of 0.01–0.02 pH units and the effect is that the quality of the pH definition – the precision by which p*K* values used in the calculations are given and the precision of experimental p*I* values in these cases – will limit the possibilities to verify polypeptide compostion based on the experimental p*I* value.

Statistical comparison of experimental and calculated p*I* values was done using the *t*-test for dependent samples and normality of the discrepancies was estimated by probability plots. For the 36 proteins, the *p*-level is 0.0021, indicating that a result like this is unlikely to be a chance effect and must be assumed to represent a real difference. After correction for the most likely *N*-terminal configuration, the *p*-level is 0.043 and cannot be accepted as representing the same population since the *p*-level is less than 0.05 – the traditional *p*-limit of statistical significance. For the 18 proteins with a known or very likely *N*-terminal configuration the *t*-test gave a *p*-level of 0.49, which verifies that the experimental and calculated p*I* values are not significantly different.

Besides showing that p*I* values for denatured proteins with known compositions can be calculated with a high degree of precision from average p*K* values, the results also provide strong support for the notion that *N*-terminal blockage heavily depends on the nature of the *N*-terminal groups [26]. The results seem to indicate that with *N*-terminals other than M, S and A, only a few proteins have blocked *N*-terminals (1 out of 10 proteins in the present study), while it can be inferred from the data presented in Table 2 that a majority of the proteins with M, S and A as *N*-terminal are blocked. After correction for the effect of suspected *N*-terminal blockage there is only one protein (nucleolar protein B23) out of the 36 used in this study, which, in spite of a high buffer capacity, has a marked difference of 0.11 pH units between predicted and determined p*I* values (Fig. 4B); this corresponds to 3 charge units due to the high buffer capacity of this protein. This discrepancy in p*I* prediction and calculation of net charge at the p*I* is probably not due to deficiencies in the database information but instead reflects a shortcoming of the model used for p*I* calculations. Nucleolar protein B23 contains a domain extremely rich in aspartic and glutamic acid residues (Table 4), in which 26 out of 28 amino acid residues from position 161 to 188 are either a D or an E. A calculation based on the use of average p*K* values uninfluenced by the charged neighboring amino acid residues cannot be expected to correctly describe the p*I* value with almost half of the acidic groups packed

Table 4. Amino acid sequence of nucleolar phosphoprotein B23

1	MEDSMDMDMS	PLRPQNYLFG	CELKADKDYH	FKVDNDENEH	QLSLRTVSLG
51	AGAKDELHIV	EAEAMNYEGS	PIKVTLATLK	MSVQPTVSLG	GFEITPPVVL
101	RLKCGSGPVH	ISGQHLVAVE	EDAESEDEEE	EDVKLLSISG	KRSAPGGGSK
151	VPQKKVKLAA	DEDDDDDDEE	DDDEDDDDDD	FDDEEAEEKA	PVKKSIRDTP
201	AKNAQKSNQN	GKDSKPSSTP	RSKGQESFKK	QEKTPKTPKG	PSSVEDIKAK
251	MQASIEKGGS	LPKVEAKFIN	YVKNCFRMTD	QEAIQDLWQW	RKSL

together into a highly negatively charged region. This limitation caused by calculations based on average p*K* values does not severely limit the usefulness of the approach since a search through Swiss-Prot shows that this type of D/E-rich motif is uncommon, and the existence of a highly charged region is immediately apparent upon inspection of the amino acid sequence.

The quality of the information available in databases, especially concerning posttranslational modifications, is a major problem when the data is to be used for p*I* predictions. The *p*-level of 0.043 found for all 36 proteins after correction for *N*-acetylation, shows that this problem is not only limited to *N*-terminal blockage and the very good agreement found for the eighteen polypeptides, with assumingly correctly described *N*-terminal (Fig. 4C), must be regarded as an exception from this point of view. *N*-Terminal blockage is generally the main problem in relation to p*I* predictions for eukaryotic proteins. Of the 36 keratinocyte proteins analyzed, 18–20 are suspected to be *N*-terminally blocked (6 proteins blocked according to Swiss-Prot, 12 proteins with M, S or A as *N*-terminal and assumingly blocked based on the calculated charge, and two proteins, involucrin and nucleolar protein B23, with M as *N*-terminal for which the data does not allow any conclusion). This is in reasonable agreement with the conclusions based on the *N*-terminal sequencing data derived in connection with 2-D gel electrophoresis. *N*-terminal blockage can be suspected for 17–19 of the 26 proteins with M, S or A as *N*-terminal, while only 1 in 10 proteins with other *N*-terminal groups are blocked. The information that the frequency of *N*-terminal blockage is strongly related to the nature of the *N*-terminal group will be of some help in connection with p*I* predictions based on database information. However, without information from other sources, an uncertainty will always remain as to whether the *N*-terminal charge should be included in the p*I* calculation.

4 Concluding remarks

The data presented here lays the foundation for comparing 2-D gel protein maps of different cell types generated with nonlinear, wide-range IPGs in the first dimension. The focusing positions of 41 polypeptides common to most human cell types have been described in a pH scale that allows focusing positions to be predicted with a high degree of accuracy, provided that the composition of the polypeptides are known and that information on posttranslational modifications are available. For polypeptides with a very high buffer capacity, the limiting factor is the precision with which experimental pH values can be determined rather than the precision of the calculations. Possible deficiencies in the pH scale description of the variation of the hydrogen ion activity has, at least at the present state, no consequences for its practical use. The major limitation in connection with predictions of focusing positions from polypeptide compositions is the quality of existing data on protein compositions, especially concerning posttranslational modifications. Amino acid sequences have been reasonably easy to obtain, while posttranslational modifications

have been difficult and work-intensive to determine. Recent developments in the field of mass spectrometry are fast changing this situation and within the next years we can expect a surge in reliable data in this area. While awaiting this development, verification of correctness and completeness of available information on polypeptide composition can be provided by experimental p*I* values in a pH scale based on the p*I* values determined in this study. So far, our data cover the pH range below pH $\approx$ 7.5. The basic pH range covered by NEPHGE as first dimension will be covered in forthcoming work.

Received December 29, 1993

5 References

[1] Gianazza, E., Astrua-Testori, S., Caccia, P., Giacon, P., Quaglia, L., Righetti, P. G., *Electrophoresis* 1986, *7*, 76–83.
[2] Görg, A., Postel, W., Günther, S., *Electrophoresis* 1988, *9*, 531–546.
[3] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.
[4] *Immobiline DryStrip Kit for 2-D Electrophoresis: Instructions*, Pharmacia LKB Biotechnology AB, Uppsala 1993.
[5] Anderson, N. L., Hickman, B. J., *Anal. Biochem.* 1979, *93*, 312–320.
[6] Neidhardt, F. C., Appleby, D. A., Sankar, P., Hutton, M. E., Phillips, T. A., *Electrophoresis* 1989, *10*, 116–121.
[7] Rasmussen, H. H., Damme, J. V., Puype, M., Gesser, B., Celis, J. E., Vandekerckhove, J., *Electrophoresis* 1992, *13*, 960–969.
[8] Gianazza, E., Artoni, G., Righetti, P. G., *Electrophoresis* 1983, *4*, 321–326.
[9] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1023–1031.
[10] Bjellqvist, B., Pasquali, C., Ravier, C., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1357–1365.
[11] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.
[12] Görg, A., *Biochem. Soc. Transactions* 1993, *21*, 130–132.
[13] Hanash, S. M., Strahler, J. R., Neel, J. V., Hailat, N., Malhem, R., Keim, D., Zhu, X. X., Wagner, D., Gage, D. A., Watson, J. T., *Proc. Natl. Acad. Sci. USA* 1991, *88*, 5709–5713.
[14] Görg, A., Postel, W., Friedrich, C., Kuick, R., Strahler, J. R., Hanash, S. M., *Electrophoresis* 1991, *12*, 653–658.
[15] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Hoffmann, H. J., Nielsen, M., Vassilev, A., Vintermyr, O., Hao, J., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjærgaard, I., Puype, M., Van Damme, J., Delay, B., Vandekerckhove, J., *Electrophoresis* 1993, *14*, 1091–1198.
[16] Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., Celis, A., Lauridsen, J. B., Basse, B., Ratz, G. P., Andersen, A., Walbum, E., Brandstrup, B., Pedersen, P. S., Brandt, N. J., Puype, M., Van Damme, J., Vandekerckhove, J., *Electrophoresis* 1991, *11*, 802–872.
[17] Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Postel, W., Westermeier, R., *J. Biochem. Biophys. Methods* 1982, *6*, 317–333.
[18] Bairoch, A., Boeckman, B., *Nucleic Acids Res.* 1991, *19*, 2247–2249.
[19] Honoré, B., Madsen, P., Basse, B., Andersen, A., Walbum, E., Celis, J. E., Leffers, H., *Nucleic Acids Res.* 1990, *18*, 6692.
[20] Altland, K., *Electrophoresis* 1990, *11*, 140–147.
[21] Perrin, D. D., Dempsey, B., Serjant, E. P., *pKa Predictions for Organic Acids and Bases*, Chapman and Hall Ltd., London 1981.
[22] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Butterworths, London 1965.
[23] Perrin, D. D., *Dissociation Constants of Organic Bases in Aqueous Solutions*, Supplement 1972, Butterworths, London 1972.
[24] Altland, K., Becher, P., Rossman, U., Bjellqvist, B., *Electrophoresis* 1988, *9*, 474–485.
[25] Brown, J. L., Robert, W. K., *J. Biol. Chem.* 1976, *251*, 1009–1014.
[26] Persson, B., Flinta, C., Heine, G., Jörnvall, H., *Eur. J. Biochem.* 1985, *152*, 523–527.

Company Info

LSB & LSP Information

Large Scale Biology Corporation
Large Scale Proteomics Corporation

Large Scale Biology Corporation

Large Scale Biology Corporation is the leader in the integrated discovery, production and application of proteins - the functional units of all biological processes.

Large Scale Biology Corporation (LSB, Vacaville, CA) and its subsidiary Large Scale Proteomics Corp. (LSP, Germantown, MD) are a biotechnology enterprise with the mission of accelerating the speed and productivity of the life sciences industry product discovery and development programs. Unique among biotechnology companies is LSB's integration of technologies to discover, analyze, manufacture and find new applications for proteins - the functional units of all biological processes.

Genomics companies have focused on deciphering genetic information, providing an initial but only partial understanding of biological processes. LSB's proprietary protein technologies can enable the transformation of genomic information into products such as drug targets, therapeutics, diagnostics for drug efficacy and toxicity, and traits for agricultural crops. Large Scale Biology has gone beyond the "genomics" realm in its business model and developed ways to integrate the discovery of gene function with quantitative protein analysis and protein manufacturing. This integration of technology platforms favorably positions LSB as a leading provider of valuable content to industry leaders in the fields of diagnostics, therapeutics, vaccines and agribusiness.

LSB was founded in 1987 with the goal of commercializing its proprietary GENEWARE viral vector system - a novel technology for gene expression. Using safe RNA viruses to transiently express genes in non-recombinant plants, LSB has positioned itself in the industry to provide cost-effective manufacturing and purification of diverse protein and peptide products. The same technology can be applied to the expression of libraries of foreign genes in an automated, high-throughput format to discover the function of genes with unparalleled efficiency. The GENEWARE system and associated proprietary technologies form the basis for LSB's functional genomics, biomanufacturing and a variety of proprietary products under development.

From its foundation, LSB understood the need to integrate functional genomic and protein manufacturing expertise with quantitative protein analysis and informatics to become a world-leader in the protein field. In 1999, LSB acquired a privately held pharmaceutical proteomics company originally founded in 1985. Large Scale Proteomics Corporation (a wholly

owned subsidiary of Large Scale Biology Corporation) is an industry leader in identifying and characterizing proteins in all types of biological samples for the discovery and development of new and more effective therapies, diagnostics, and agricultural products.

"Proteomics" is the study of the entire complement of proteins expressed in a cell, tissue, or organism. Proteomics can significantly improve drug discovery and development because most illness is associated with imbalances among, or malfunctions of, proteins. Only a small fraction of diseases can be attributed to the presence of a defective gene. Unlike classical genomics approaches that discover genes that may relate to a disease, LSP has developed a proprietary system called the ProGEx module for directly characterizing proteins associated with disease. Using this same technology, LSP can characterize the effects of candidate drugs intended to reverse a disease process, and to determine the degree to which this objective is achieved free of adverse side effects.

LSB and LSP have protected their many discoveries though an extensive portfolio of domestic and foreign patents and have developed commercial alliances and partnerships to exploit the value of their technologies. LSB and LSP scientists and engineers focus on the development and application of resources to help clients meet their objectives as well as the development of our own proprietary products for subsequent partnering with industry leaders.

A combined staff of 140 professionals operates from three locations in the United States, with a network of collaborators and affiliates throughout the US and Europe. Company headquarters, R&D laboratories and its Genomics division are located in Vacaville, California about 60 miles northeast of San Francisco. Process development and biomanufacturing take place in Owensboro, Kentucky, and LSB's Large Scale Proteomics Corporation subsidiary is located in Germantown, Maryland.

In August, 2000, LSB completed an initial public offering (IPO) of 5 million shares of common stock and now trades on the NASDAQ under the symbol LSBC.

Leadership - Large Scale Biology Corporation

Robert L. Erwin, Chairman of the Board and Chief Executive Officer, founded LSB™ and has served as a director and officer since 1987. Mr. Erwin is the former chairman of the State of California Breast Cancer Research Council and currently serves on the University of California President's Engineering Advisory Council. He is Chairman of the Supervisory Board of Icon Genetics AG. As a co-founder of Sungene Technologies Corp., Mr. Erwin served as Vice President of Research and Product Development from 1981 through 1986. He has served on the Biotechnology Industry Advisory Board for Iowa State University. Mr. Erwin received his M.S. degree in Genetics from Louisiana State University and is an inventor on several LSB patents.

*David R. McGee, Ph.D.,*a co-founder of LSB and Senior Vice President and Chief Operating Officer, has been an officer since 1987. Prior to joining LSB, Dr. McGee was Vice President of Operations at Sungene Technologies Corporation from 1983 to 1987. Dr. McGee received his Ph.D. in Genetics from Louisiana State University and served as a faculty instructor of zoology and genetics at Louisiana State University.

*Laurence K. Grill, Ph.D.,*a co-founder of LSB and Senior Vice President, Research and Development, has served as an officer since 1987. Dr. Grill was the Manager of Plant Molecular Biology for Sandoz Crop Protection Corp. from 1984 to 1987 and Senior Research

Scientist in the Department of Molecular Biology at Zoecon Research Institute from 1980 to 1984. He received his Ph.D. from the University of California at Riverside with an emphasis on the molecular basis for viral gene expression in plants.

R. Barry Holtz, Ph. D., Senior Vice President, Biopharmaceutical Manufacturing, has served the company as an officer since 1989 upon the acquisition of Holtz Bio-Engineering, which was founded in 1980. Dr. Holtz was a co-founder and Director of Research for MFI, Inc., the largest manufacturer of microencapsulated nutrients for agriculture and Director of Fundamental Research at Foremost-McKesson, Inc. Dr. Holtz received his Ph.D. in Biochemistry from Pennsylvania State University and served as Assistant Professor in the Department of Food Science and Nutrition at Ohio State University.

Daniel Tusé, Ph.D., has been an officer of LSB since he joined the Company in 1995 as Vice President, Pharmaceutical Development. Dr. Tusé manages the company's pharmaceutical design and development programs, including LSB's novel vaccines and immunotherapeutics initiatives. Prior to joining LSB, Dr. Tusé was Assistant Director of SRI International's (Menlo Park, Calif.) Life Sciences Division. In his 17 years at SRI, Dr. Tusé developed extensive R&D experience in pharmaceuticals and specialty chemicals, serving an international list of clients. Dr. Tusé received his Ph.D. in Microbiology (1980, *cum laude*) with a minor in Toxicology from the University of California, Davis.

John S. Rakitan, a co-founder of LSB, Senior Vice President & General Counsel and Secretary, has served as an officer since 1988. Prior to joining LSB, Mr. Rakitan was an attorney in private practice. Mr. Rakitan received his J.D. degree from the University of Notre Dame.

Michael D. Centron, Treasurer, has served as Controller since 1988 and was elected as Treasurer in 1991. Mr. Centron was Audit Supervisor for Varian Associates from June 1985 through July 1988, and he also worked for Arthur Young and Co. (currently Ernst & Young). Mr. Centron is a certified public accountant and received his M.B.A. degree from the University of California at Berkeley.

Guy della-Cioppa, Ph.D., is an officer of the company and currently serves as Vice President, Genomics. Prior to joining the company in 1989, Dr. della-Cioppa worked for Monsanto Company in St. Louis, MO from 1984-1989 and was an NIH Postdoctoral Fellow at the Worcester Foundation for Experimental Biology in Shrewsbury, MA from 1983-1984. He received his Ph.D. in Biology from the University of California, Los Angeles.

William M. Pfann joined Large Scale Biology in August 2000 as Senior Vice President Finance and Chief Financial Officer. Mr. Pfann was formerly with PricewaterhouseCoopers LLP from 1969 to July 2000, most recently as the Risk Management Partner for the Western Region. He served in a number of management roles at PwC, including leader of the firm's Silicon Valley audit practice, National Director of the networking and communications sector and Managing Partner of the Northern California emerging business group, as well as Partner-in-Charge of the Oakland and Walnut Creek, California offices. Mr. Pfann received a B.S. degree from the University of California, Berkeley, in Business Administration and an MBA in Accounting from Golden Gate University.

back to index

Large Scale Proteomics Corporation

Leadership - Large Scale Proteomics Corporation

*N. Leigh Anderson, Ph D.,*Chairman, President and CEO of Large Scale Proteomics Corporation (LSP™). Dr. Anderson obtained his B.A. in Physics with honors from Yale and a Ph.D. in Molecular Biology from Cambridge University (England) working with M. F. Perutz as a Churchill Fellow at the MRC Laboratory of Molecular Biology. Subsequently he co-founded the Molecular Anatomy Program at the Argonne National Laboratory (Chicago) where his work in the development of 2-dimensional electrophoresis (2-DE) and molecular database technology earned him, among other distinctions, the American Association for Clinical Chemistry's Young Investigator Award for 1982 and the 1983 Pittsburgh Analytical Chemistry Award. In 1985 Dr. Anderson co-founded LSP (originally Large Scale Biology Corp., Germantown, MD) in order to pursue commercial development and large-scale applications of 2-D electrophoretic protein mapping technology.

*Norman G. Anderson, Ph.D.,*Chief Scientist at LSP. Dr. Anderson has a distinguished record as an inventor. His career includes senior positions at Oak Ridge and Argonne National Laboratories (ORNL and ANL), more than 300 scientific publications, and the receipt of more than 20 prestigious awards in recognition of his work in science and technology. For his invention of the zonal ultracentrifuge, he received the John Scott Medal Award, and for the centrifugal fast analyzer, the Preis Biochemische Analytik für Klinische Chemie from Die Deutsche Gesellschaft für Klinische Chemie for the most outstanding analytical development in clinical chemistry worldwide during a 2-year period. In 1984 ANL awarded him its career patent leader award for the largest number of patents issued to an employee. At that time the commercial value of his inventions in terms of U.S. sales and royalties from foreign licensing were $250 million and $1 million, respectively. Dr. Anderson received his degrees at Duke University: a B.A. in Zoology, M.A. in Physiology, and Ph.D. in Cell Physiology. He holds 28 patents.

*Constance Seniff,*Vice President, Operations. Ms. Seniff has managed LSP's operations since 1993. Her background includes thirteen years in international business prior to joining LSP, five abroad in the employ of foreign firms. Ms. Seniff is responsible for helping formulate and implement business development and database commercialization strategies for LSP in coordination with the management of LSP's parent company, Large Scale Biology Corporation. Ms. Seniff has a B.Sc. degree in Business (with honors) from Florida State University.

Robert J. Walden, Vice President, Finance at LSP. Mr. Walden joined LSP in 1997 and has served as a director since 1999. He previously served as Vice President of Finance and Administration at Osiris Therapeutics, Inc., and as Chief Financial Officer at the American Type Culture Collection (ATCC). Mr. Walden received his degree in Finance from the University of Maryland.

*Jean-Paul Hofmann, Ph.D.,*Vice President, Software Development at LSP. Dr. Hofmann is a plant geneticist by training, having earned a B.S. in Biology, M.S. in Biochemistry and Genetics, and Ph.D. in Plant Genetics from the University of Orsay, Paris. He has extensive

experience in using 2-DE in agronomic research and in designing analytical software for 1- and 2-D applications. He has held senior scientific positions in industry and research institutes, in the U.S., France and the Ivory Coast.

*John Taylor, Ph.D.,*Vice President, Software Development and Bioinformatics. Dr. Taylor is the principal developer of Kepler™, LSP's analytical software for automated 2-DE pattern analysis. Prior to joining LSB, Dr. Taylor served as computer scientist in the Molecular Anatomy Program at Argonne, and on the research staffs of the University of Chicago and the Armed Forces Institute of Pathology in Washington, D.C. Dr. Taylor received a B.S. in Physics from the University of South Carolina, and a Ph.D. in Nuclear Physics from Duke University.

Sandra Steiner, Ph.D., currently serves as Vice President Proteomics Applications. Prior to joining the Company, Dr. Steiner founded and directed the Molecular Toxicology Group at Novartis in Basel, Switzerland and was a member in several multi-disciplinary drug development project teams. Dr. Steiner received her Ph.D. in Toxicology/Pharmacology from the University of Basel, Switzerland.

back to index

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Exhib[illegible]

SeqServer®
biology in silico

BLAST2 Search Results

Sequences | Help

Retrieval | BLAST2 | FASTA | ClustalW | GCG Assembly | Phrap | Translation | BLAST2 Manual

Confidential – Property of Incyte Genomics, Inc. SeqServer Version 4.6 Jan 2002

```
Program: blastp
Sequence ID(s):
☐ PF-0232-1DIV_HNT2NOT01 vs. genpept132

NCBI-BLASTP 2.0.10 [Aug-26-1999]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= PF-0232-1DIV_HNT2NOT01
         (491 letters)

Database: genpept132
           1,206,111 sequences; 372,583,108 total letters

Searching..................................................done

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

☑ g21618457  tubby like protein 3 [Homo sapiens]                  886     0.0
☑ g3372493  tubby like protein 3 [Homo sapiens]                   825     0.0
☑ g4239948  tubby [Mus musculus]                                  628   e-179
☑ g3372491  tubby like protein 3 [Mus musculus]                   628   e-179
☑ g3551050  TUBBY protein [Rattus norvegicus]                     454   e-126
☑ g1305497  tub homolog [Homo sapiens]                            453   e-126
☑ g2072160  tub homolog [Homo sapiens]                            451   e-125
☑ g14253111  tubby protein [Mus musculus]                         450   e-125
☑ g1305499  tubby [Mus musculus]                                  450   e-125
☑ g11071535  tubby (mouse) homolog [Homo sapiens]                 421   e-116

>g21618457 tubby like protein 3 [Homo sapiens]
          Length = 442

 Score =  886 bits (2264), Expect = 0.0
 Identities = 435/439 (99%), Positives = 437/439 (99%)

Query: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60
            MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA
Sbjct: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60

Query: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS 120
            KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS
Sbjct: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS 120

Query: 121  KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTGTSGSATAAQPADN 180
            KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTGTSGSATAAQPADN
Sbjct: 121  KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTGTSGSATAAQPADN 180

Query: 181  LLGDIDDLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARK 240
            LLGDID LEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARK
Sbjct: 181  LLGDIDYLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARK 240

Query: 241  RKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR 300
```



```
             RKKSKTANYLISIDPVDL[illegible]YVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR
Sbjct: 241  RKKSKTANYLISIDPVD[illegible]VGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR 300

Query: 301  QELAAISYETNVLGFKGPRK[illegible]SVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE 360
            QELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE
Sbjct: 301  QELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE 360

Query: 361  LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNY 420
            LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNY
Sbjct: 361  LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNY 420

Query: 421  PLCAVQAFGIGLSSFDKRI 439
            PLCAVQAFGIGLSSFD ++
Sbjct: 421  PLCAVQAFGIGLSSFDSKL 439

>g3372493 tubby like protein 3 [Homo sapiens]
          Length = 442

 Score =  825 bits (2109), Expect = 0.0
 Identities = 410/439 (93%), Positives = 413/439 (93%)

Query: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60
            MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA
Sbjct: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60

Query: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS 120
            KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS
Sbjct: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPDEVHAPSVSSSVVEEDAENTVDTAS 120

Query: 121  KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTGTSGSATAAQPADN 180
            KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTG
Sbjct: 121  KPGLQERLQKHDISESVNFDEETDGISQSACLERPNSASSQNSTDTGIPVLLLPPNQLIT 180

Query: 181  LLGDIDDLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARK 240
             LGDIDDLEDFV  PAPQGVTVRCRIIRDKRGMDRGLF        L KEENQKIFLLAARK
Sbjct: 181  FLGDIDDLEDFVLVPAPQGVTVRCRIIRDKRGMDRGLFSHLLYVLGKEENQKIFLLAARK 240

Query: 241  RKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR 300
            RKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR
Sbjct: 241  RKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHTR 300

Query: 301  QELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE 360
            QELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE
Sbjct: 301  QELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVE 360

Query: 361  LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNY 420
            LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQ+VHKNDPDYIVMQFGRVADDVFTLDYNY
Sbjct: 361  LHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQLVHKNDPDYIVMQFGRVADDVFTLDYNY 420

Query: 421  PLCAVQAFGIGLSSFDKRI 439
            PLCAVQAF I LSSFD ++
Sbjct: 421  PLCAVQAFAISLSSFDSKL 439

>g4239948 tubby [Mus musculus]
          Length = 460

 Score =  628 bits (1602), Expect = e-179
 Identities = 314/459 (68%), Positives = 364/459 (78%), Gaps = 22/459 (4%)

Query: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60
            MEA+RC   P GDS F +E +++RQ KLD QR LLEK+QRKKRLEP MVQPNPEARLRR
Sbjct: 1    MEAARCAPGPRGDSAFDDETLRLRQLKLDNQRALLEKKQRKKRLEPLMVQPNPEARLRRL 60

Query: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPD------EVHAPSVSSSVVEEDAEN 114
            KPR S+E TPLV+   P S+VILHGIDGPAA LKP+       +      SV S    EE  E
Sbjct: 61   KPRGSEEHTPLVDPQMPRSDVILHGIDGPAAFLKPEAQDLESKPQVLSVGSPAPEEGTEG 120

Query: 115  TVD------TASKPGLQERLQKHDISESVNFDEETDG--------ISQSACLERPNSASS 160
            + D      TA KP LQE LQKH I  SVN+DEE D         S SA  E  +AS
Sbjct: 121  SADGESPEETAPKPDLQEILQKHGILSSVNYDEEPDKEEDEGGNLSSPSARSEESAAASQ 180

Query: 161  QNSTDTGTSGSATAAQPADNLLGDIDDLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPT 220
```



```
            + +++TG SG    AQ' D  LG++++LEDF YSPAP+GVTV+C++ RDK+GMDRGLFPT
Sbjct: 181  KAASETGASG--VTAQQGDAQLGEVENLEDFAYSPAPRGVTVKCKVTRDKKGMDRGLFPT 238

Query: 221  YYMYLEKEENQKIFLLAARKRKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTV 280
            YYM+LE+EEN+KIFLLA RKRKKSKT+NYL+S DP DLSREGESY+GKLRSNLMGTKFTV
Sbjct: 239  YYMHLEREENRKIFLLAGRKRKKSKTSNYLVSTDPTDLSREGESYIGKLRSNLMGTKFTV 298

Query: 281  YDRGICPMKGRGLVGAAHTRQELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQ 340
            YD G+ P+K +GLV  AHTRQELAAI YETNVLGFKGPRKMSVIIPGM +NH++IP++P+
Sbjct: 299  YDHGVNPVKAQGLVEKAHTRQELAAICYETNVLGFKGPRKMSVIIPGMNMNHERIPFRPR 358

Query: 341  NNHDSLLSRWQNRTMENLVELHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPD 400
            N H+SLLS+WQN++MENL+ELHNKAPVWN DTQSYVLNF GRVTQASVKNFQIVH NDPD
Sbjct: 359  NEHESLLSKWQNKSMENLIELHNKAPVWNDDTQSYVLNFHGRVTQASVKNFQIVHGNDPD 418

Query: 401  YIVMQFGRVADDVFTLDYNYPLCAVQAFGIGLSSFDKRI 439
            YIVMQFGRVADDVFTLDYNYPLCA+QAF IGLSSFD ++
Sbjct: 419  YIVMQFGRVADDVFTLDYNYPLCALQAFAIGLSSFDSKL 457

>g3372491 tubby like protein 3 [Mus musculus]
          Length = 460

 Score =  628 bits (1602), Expect = e-179
 Identities = 314/459 (68%), Positives = 364/459 (78%), Gaps = 22/459 (4%)

Query: 1    MEASRCRLSPSGDSVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRA 60
            MEA+RC   P GDS F +E +++RQ KLD QR LLEK+QRKKRLEP MVQPNPEARLRR
Sbjct: 1    MEAARCAPGPRGDSAFDDETLRLRQLKLDNQRALLEKKQRKKRLEPLMVQPNPEARLRRL 60

Query: 61   KPRASDEQTPLVNCHTPHSNVILHGIDGPAAVLKPD------EVHAPSVSSSVVEEDAEN 114
            KPR S+E TPLV+   P S+VILHGIDGPAA LKP+       +      SV S    EE  E
Sbjct: 61   KPRGSEEHTPLVDPQMPRSDVILHGIDGPAAFLKPEAQDLESKPQVLSVGSPAPEEGTEG 120

Query: 115  TVD------TASKPGLQERLQKHDISESVNFDEETDG--------ISQSACLERPNSASS 160
            + D      TA KP LQE LQKH I  SVN+DEE D         S SA  E  +AS
Sbjct: 121  SADGESPEETAPKPDLQEILQKHGILSSVNYDEEPDKEEDEGGNLSSPSARSEESAAASQ 180

Query: 161  QNSTDTGTSGSATAAQPADNLLGDIDDLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPT 220
            + +++TG SG    AQ  D  LG++++LEDF YSPAP+GVTV+C++ RDK+GMDRGLFPT
Sbjct: 181  KAASETGASG--VTAQQGDAQLGEVENLEDFAYSPAPRGVTVKCKVTRDKKGMDRGLFPT 238

Query: 221  YYMYLEKEENQKIFLLAARKRKKSKTANYLISIDPVDLSREGESYVGKLRSNLMGTKFTV 280
            YYM+LE+EEN+KIFLLA RKRKKSKT+NYL+S DP DLSREGESY+GKLRSNLMGTKFTV
Sbjct: 239  YYMHLEREENRKIFLLAGRKRKKSKTSNYLVSTDPTDLSREGESYIGKLRSNLMGTKFTV 298

Query: 281  YDRGICPMKGRGLVGAAHTRQELAAISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQ 340
            YD G+ P+K +GLV  AHTRQELAAI YETNVLGFKGPRKMSVIIPGM +NH++IP++P+
Sbjct: 299  YDHGVNPVKAQGLVEKAHTRQELAAICYETNVLGFKGPRKMSVIIPGMNMNHERIPFRPR 358

Query: 341  NNHDSLLSRWQNRTMENLVELHNKAPVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPD 400
            N H+SLLS+WQN++MENL+ELHNKAPVWN DTQSYVLNF GRVTQASVKNFQIVH NDPD
Sbjct: 359  NEHESLLSKWQNKSMENLIELHNKAPVWNDDTQSYVLNFHGRVTQASVKNFQIVHGNDPD 418

Query: 401  YIVMQFGRVADDVFTLDYNYPLCAVQAFGIGLSSFDKRI 439
            YIVMQFGRVADDVFTLDYNYPLCA+QAF IGLSSFD ++
Sbjct: 419  YIVMQFGRVADDVFTLDYNYPLCALQAFAIGLSSFDSKL 457

>g3551050 TUBBY protein [Rattus norvegicus]
          Length = 505

 Score =  454 bits (1156), Expect = e-126
 Identities = 244/491 (49%), Positives = 316/491 (63%), Gaps = 66/491 (13%)

Query: 14   SVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRAKPRASDEQTPLVN 73
            SV  +E   +RQ KLD QR LLE++Q+KKR EP MVQ N + R R  + R S+EQ PLV
Sbjct: 13   SVLDDEGSNLRQQKLDRQRALLEQKQKKKRQEPLMVQANADGRPRSRRARQSEEQAPLVE 72

Query: 74   CHTPHSNVILH-------------------------------------------------- 84
             +   S   +
Sbjct: 73   SYLSSSGSTSYQVQEADSLASVQPGATRPPAPASAKKTKGAAASGGQGGAPRKEKKGKHK 132

Query: 85   GIDGPAAVLKP-DEVHAPSVSSSVVEED-AENTVDTASKPG-------LQERLQKHDISE 135
```



```
            G  GPA + +    E    P     +V + D A++  +TA+  G          L+  +Q+  IS
Sbjct: 133  GTSGPATLAEDKSEAQGPVQILTVGQSDHAKDAGETAAGGGAQPSGQDLRATMQRKGISS 192

Query: 136  SVNFDEETD------GISQSACLERPNSASSQNSTDTGTSGSATAAQPADNLLGDIDDLE 189
            S++FDEE D         SQ       RP+SA+S+ ST    S  + AA P   +  ++ DLE
Sbjct: 193  SMSFDEEEDEDENSSSSSQLNSNTRPSSATSRKSTREAASAPSPAA-PEPPVDIEVQDLE 251

Query: 190  DFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKTANY 249
            +F   PAPQG+T++CRI RDK+GMDRG++PTY++L++E+ +K+FLLA RKRKKSKT+NY
Sbjct: 252  EFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKTSNY 311

Query: 250  LISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAAISY 308
            LIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K          + T RQELAA+ Y
Sbjct: 312  LISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAAVCY 371

Query: 309  ETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKAPVW 368
            ETNVLGFKGPRKMSVI+PGM + H+++ +P+N H++LL+RWQN+ E++EL NK PVW
Sbjct: 372  ETNVLGFKGPRKMSVIVPGMNMVHERVCIRPRNEHETLLARWQNKNTESIIELQNKTPVW 431

Query: 369  NSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNYPLCAVQAF 428
            N DTQSYVLNF GRVTQASVKNFQI+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+QAF
Sbjct: 432  NDDTQSYVLNFHGRVTQASVKNFQIIHGNDPDYIVMQFGRVAEDVFTMDYNYPLCALQAF 491

Query: 429  GIGLSSFDKRI 439
             I LSSFD ++
Sbjct: 492  AIALSSFDSKL 502

>g1305497 tub homolog [Homo sapiens]
          Length = 506

 Score =  453 bits (1152), Expect = e-126
 Identities = 245/494 (49%), Positives = 318/494 (63%), Gaps = 71/494 (14%)

Query: 14   SVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRAKPRASDEQTPLVN 73
            SV  +E   +RQ KLD QR LLE++Q+KKR EP MVQ N + R R  + R S+EQ PLV
Sbjct: 13   SVLDDEGRNLRQQKLDRQRALLEQKQKKKRQEPLMVQANADGRPRSRRARQSEEQAPLVE 72

Query: 74   CHTPHSNVILH-------------------------------------------------- 84
             +   S   +
Sbjct: 73   SYLSSSGSTSYQVQEADSLASVQLGATRPTAPASAKRTKAAATAGGQGGAARKEKKGKHK 132

Query: 85   GIDGPAAVLKP-DEVHAPSVSSSVVEED-AENTVDTASKPG-------LQERLQKHDISE 135
            G  GPAA+ +    E    P     +V + D A++  +TA+  G          L+  +Q+  IS
Sbjct: 133  GTSGPAALAEDKSEAQGPVQILTVGQSDHAQDAGETAAGGGERPSGQDLRATMQRKGISS 192

Query: 136  SVNFDE----ETDGISQSACLE---RPNSASSQNSTDTGTSGSATAA--QPADNLLGDID 186
            S++FDE    E +  S S+ L     RP+SA+S+ S     S  + A  QP D     ++
Sbjct: 193  SMSFDEDEEDEEENSSSSSQLNSNTRPSSATSRKSVREAASAPSPTAPEQPVDV---EVQ 249

Query: 187  DLEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKT 246
            DLE+F   PAPQG+T++CRI RDK+GMDRG++PTY++L++E+ +K+FLLA RKRKKSKT
Sbjct: 250  DLEEFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKT 309

Query: 247  ANYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAA 305
            +NYLIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K          + T RQELAA
Sbjct: 310  SNYLISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAA 369

Query: 306  ISYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKA 365
            + YETNVLGFKGPRKMSVI+PGM + H+++ +P+N H++LL+RWQN+ E++EL NK
Sbjct: 370  VCYETNVLGFKGPRKMSVIVPGMNMVHERVSIRPRNEHETLLARWQNKNTESIIELQNKT 429

Query: 366  PVWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNYPLCAV 425
            PVWN DTQSYVLNF GRVTQASVKNFQI+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+
Sbjct: 430  PVWNDDTQSYVLNFHGRVTQASVKNFQIIHGNDPDYIVMQFGRVAEDVFTMDYNYPLCAL 489

Query: 426  QAFGIGLSSFDKRI 439
            QAF I LSSFD ++
Sbjct: 490  QAFAIALSSFDSKL 503

>g2072160 tub homolog [Homo sapiens]
          Length = 561
```

```
 Score =  451 bits (1148), Expect = e-125
 Identities = 244/493 (49%), Positives = 317/493 (63%), Gaps = 7[illegible]14%)

Query: 15   VFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRAKPRASDEQT[illegible]LVNC 74
            V  +E   +RQ KLD QR LLE++Q+KKR EP MVQ N + R R  + R S+EQ PLV
Sbjct: 69   VLDDEGRNLRQQKLDRQRALLEQKQKKKRQEPLMVQANADGRPRSRRARQSEEQAPLVES 128

Query: 75   HTPHSNVILH-------------------------------------------------G 85
            +   S    +                                                 G
Sbjct: 129  YLSSSGSTSYQVQEADSLASVQLGATRPTAPASAKRTKAAATAGGQGGAARKEKKGKHKG 188

Query: 86   IDGPAAVLKP-DEVHAPSVSSSVVEED-AENTVDTASKPG-------LQERLQKHDISES 136
              GPAA+ +    E   P   +V + D A++  +TA+  G       L+  +Q+  IS S
Sbjct: 189  TSGPAALAEDKSEAQGPVQILTVGQSDHAQDAGETAAGGGERPSGQDLRATMQRKGISSS 248

Query: 137  VNFDE----ETDGISQSACLE---RPNSASSQNSTDTGTSGSATAA--QPADNLLGDIDD 187
            ++FDE    E +  S S+ L    RP+SA+S+ S      S  +  A  QP D   ++ D
Sbjct: 249  MSFDEDEEDEEENSSSSSQLNSNTRPSSATSRKSVREAASAPSPTAPEQPVDV---EVQD 305

Query: 188  LEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKTA 247
            LE+F    PAPQG+T++CRI RDK+GMDRG++PTY+++L++E+ +K+FLLA RKRKKSKT+
Sbjct: 306  LEEFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKTS 365

Query: 248  NYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAAI 306
            NYLIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K        + T RQELAA+
Sbjct: 366  NYLISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAAV 425

Query: 307  SYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKAP 366
             YETNVLGFKGPRKMSVI+PGM + H+++   +P+N H++LL+RWQN+  E++EL NK P
Sbjct: 426  CYETNVLGFKGPRKMSVIVPGMNMVHERVSIRPRNEHETLLARWQNKNTESIIELQNKTP 485

Query: 367  VWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNYPLCAVQ 426
            VWN DTQSYVLNF GRVTQASVKNFQI+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+Q
Sbjct: 486  VWNDDTQSYVLNFHGRVTQASVKNFQIIHGNDPDYIVMQFGRVAEDVFTMDYNYPLCALQ 545

Query: 427  AFGIGLSSFDKRI 439
            AF I LSSFD ++
Sbjct: 546  AFAIALSSFDSKL 558


>g14253111 tubby protein [Mus musculus]
          Length = 505

 Score =  450 bits (1144), Expect = e-125
 Identities = 242/491 (49%), Positives = 314/491 (63%), Gaps = 66/491 (13%)

Query: 14   SVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRAKPRASDEQTPLVN 73
            SV  +E   +RQ KLD QR LLE++Q+KKR EP MVQ N + R R  + R S+EQ PLV
Sbjct: 13   SVLDDEGSNLRQQKLDRQRALLEQKQKKKRQEPLMVQANADGRPRSRRARQSEEQAPLVE 72

Query: 74   CHTPHSNVILH------------------------------------------------- 84
              +   S    +
Sbjct: 73   SYLSSSGSTSYQVQEADSIASVQLGATRPPAPASAKKSKGAAASGGQGGAPRKEKKGKHK 132

Query: 85   GIDGPAAVLKP-DEVHAP----SVSSSVVEEDA-ENTVDTASKPG---LQERLQKHDISE 135
            G  GPA + +    E   P    +V  S ++DA E          ++P    L+  +Q+  IS
Sbjct: 133  GTSGPATLAEDKSEAQGPVQILTVGQSDHDKDAGETAAGGAQPSGQDLRATMQRKGISS 192

Query: 136  SVNFDEETD------GISQSACLERPNSASSQNSTDTGTSGSATAAQPADNLLGDIDDLE 189
            S++FDE+ D          SQ     RP+SA+S+ S      S  + AA P    +  ++ DLE
Sbjct: 193  SMSFDEDEDEDENSSSSSQLNSNTRPSSATSRKSIREAASAPSPAA-PEPPVDIEVQDLE 251

Query: 190  DFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKTANY 249
            +F    PAPQG+T++CRI RDK+GMDRG++PTY+++L++E+ +K+FLLA RKRKKSKT+NY
Sbjct: 252  EFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKTSNY 311

Query: 250  LISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAAISY 308
            LIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K        + T RQELAA+ Y
Sbjct: 312  LISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAAVCY 371

Query: 309  ETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKAPVW 368
            ETNVLGFKGPRKMSVI+PGM + H+++   +P+N H++LL+RWQN+  E++EL NK PVW
Sbjct: 372  ETNVLGFKGPRKMSVIVPGMNMVHERVCIRPRNEHETLLARWQNKNTESIIELQNKTPVW 431
```



```
Query: 369  NSDTQSYVLNFRGRVTQAS[illegible]VHKNDPDYIVMQFGRVADDVFTLDYNYPLCAVQAF 428
            N DTQSYVLNF GRVTQA[illegible]+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+QAF
Sbjct: 432  NDDTQSYVLNFHGRVTQAS[illegible]IHGNDPDYIVMQFGRVAEDVFTMDYNYPLCALQAF 491

Query: 429  GIGLSSFDKRI 439
             I LSSFD ++
Sbjct: 492  AIALSSFDSKL 502


>g1305499 tubby [Mus musculus]
          Length = 505

 Score =  450 bits (1144), Expect = e-125
 Identities = 242/491 (49%), Positives = 314/491 (63%), Gaps = 66/491 (13%)

Query: 14   SVFHEEMMKMRQAKLDYQRLLLEKRQRKKRLEPFMVQPNPEARLRRAKPRASDEQTPLVN 73
            SV  +E   +RQ KLD QR LLE++Q+KKR EP MVQ N + R R  + R S+EQ PLV
Sbjct: 13   SVLDDEGSNLRQQKLDRQRALLEQKQKKKRQEPLMVQANADGRPRSRRARQSEEQAPLVE 72

Query: 74   CHTPHSNVILH------------------------------------------------- 84
              +   S    +
Sbjct: 73   SYLSSSGSTSYQVQEADSIASVQLGATRPPAPASAKKSKGAAASGGQGGAPRKEKKGKHK 132

Query: 85   GIDGPAAVLKP-DEVHAP----SVSSSVVEEDA-ENTVDTASKPG---LQERLQKHDISE 135
            G  GPA + +    E   P    +V  S ++DA E          ++P    L+  +Q+  IS
Sbjct: 133  GTSGPATLAEDKSEAQGPVQILTVGQSDHDKDAGETAAGGAQPSGQDLRATMQRKGISS 192

Query: 136  SVNFDEETD------GISQSACLERPNSASSQNSTDTGTSGSATAAQPADNLLGDIDDLE 189
            S++FDE+ D          SQ     RP+SA+S+ S      S  + AA P    +  ++ DLE
Sbjct: 193  SMSFDEDEDEDENSSSSSQLNSNTRPSSATSRKSIREAASAPSPAA-PEPPVDIEVQDLE 251

Query: 190  DFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKTANY 249
            +F    PAPQG+T++CRI RDK+GMDRG++PTY+++L++E+ +K+FLLA RKRKKSKT+NY
Sbjct: 252  EFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKTSNY 311

Query: 250  LISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAAISY 308
            LIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K        + T RQELAA+ Y
Sbjct: 312  LISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAAVCY 371

Query: 309  ETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKAPVW 368
            ETNVLGFKGPRKMSVI+PGM + H+++   +P+N H++LL+RWQN+  E++EL NK PVW
Sbjct: 372  ETNVLGFKGPRKMSVIVPGMNMVHERVCIRPRNEHETLLARWQNKNTESIIELQNKTPVW 431

Query: 369  NSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNYPLCAVQAF 428
            N DTQSYVLNF GRVTQASVKNFQI+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+QAF
Sbjct: 432  NDDTQSYVLNFHGRVTQASVKNFQIIHGNDPDYIVMQFGRVAEDVFTMDYNYPLCALQAF 491

Query: 429  GIGLSSFDKRI 439
             I LSSFD ++
Sbjct: 492  AIALSSFDSKL 502


>g11071535 tubby (mouse) homolog [Homo sapiens]
          Length = 420

 Score =  421 bits (1071), Expect = e-116
 Identities = 213/373 (57%), Positives = 278/373 (74%), Gaps = 21/373 (5%)

Query: 85   GIDGPAAVLKP-DEVHAPSVSSSVVEED-AENTVDTASKPG-------LQERLQKHDISE 135
            G  GPAA+ +    E   P   +V + D A++  +TA+  G       L+  +Q+  IS
Sbjct: 48   GTSGPAALAEDKSEAQGPVQILTVGQSDHAQDAGETAAGGGERPSGQDLRATMQRKGISS 107

Query: 136  SVNFDE----ETDGISQSACLE---RPNSASSQNSTDTGTSGSATAA-QPADNLLGDIDD 187
            S++FDE    E +  S S+ L    RP+SA+S+ S +  ++ S TA  QP D   ++ D
Sbjct: 108  SMSFDEDEEDEEENSSSSSQLNSNTRPSSATSRKSVEAASAPSPTAPEQPVDV---EVQD 164

Query: 188  LEDFVYSPAPQGVTVRCRIIRDKRGMDRGLFPTYYMYLEKEENQKIFLLAARKRKKSKTA 247
            LE+F    PAPQG+T++CRI RDK+GMDRG++PTY+++L++E+ +K+FLLA RKRKKSKT+
Sbjct: 165  LEEFALRPAPQGITIKCRITRDKKGMDRGMYPTYFLHLDREDGKKVFLLAGRKRKKSKTS 224

Query: 248  NYLISIDPVDLSREGESYVGKLRSNLMGTKFTVYDRGICPMKGRGLVGAAHT-RQELAAI 306
            NYLIS+DP DLSR G+SY+GKLRSNLMGTKFTVYD G+ P K        + T RQELAA+
Sbjct: 225  NYLISVDPTDLSRGGDSYIGKLRSNLMGTKFTVYDNGVNPQKASSSTLESGTLRQELAAV 284
```



```
Query: 307  SYETNVLGFKGPRKMSVIIPGMTLNHKQIPYQPQNNHDSLLSRWQNRTMENLVELHNKAP 366
             YETNVLGFKGPRKMSVI+PGM + H+++   +P+N H++LL+RWQN+  E++EL NK P
Sbjct: 285  CYETNVLGFKGPRKMSVIVPGMNMVHERVSIRPRNEHETLLARWQNKNTESIIELQNKTP 344

Query: 367  VWNSDTQSYVLNFRGRVTQASVKNFQIVHKNDPDYIVMQFGRVADDVFTLDYNYPLCAVQ 426
            VWN DTQSYVLNF GRVTQASVKNFQI+H NDPDYIVMQFGRVA+DVFT+DYNYPLCA+Q
Sbjct: 345  VWNDDTQSYVLNFHGRVTQASVKNFQIIHGNDPDYIVMQFGRVAEDVFTMDYNYPLCALQ 404

Query: 427  AFGIGLSSFDKRI 439
            AF I LSSFD ++
Sbjct: 405  AFAIALSSFDSKL 417


  Database: genpept132
    Posted date:  Nov 14, 2002  2:57 PM
  Number of letters in database: 372,583,108
  Number of sequences in database:  1,206,111

Lambda     K      H
   0.317    0.133    0.390

Gapped
Lambda     K      H
   0.270   0.0470    0.230


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 362253949
Number of Sequences: 1206111
Number of extensions: 15042333
Number of successful extensions: 41824
Number of sequences better than 10.0: 118
Number of HSP's better than 10.0 without gapping: 75
Number of HSP's successfully gapped in prelim test: 43
Number of HSP's that attempted gapping in prelim test: 41485
Number of HSP's gapped (non-prelim): 225
length of query: 491
length of database: 372,583,108
effective HSP length: 60
effective length of query: 431
effective length of database: 300,216,448
effective search space: 129393289088
effective search space used: 129393289088
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.7 bits)
```

Graphical Viewer

Submit sequences to: BLAST2 Submit Reset

IncyteGenomics

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref. No. 2

MOLECULAR CARCINOGENESIS 24:153–159 (1999)

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1]*Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*

[2]*Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

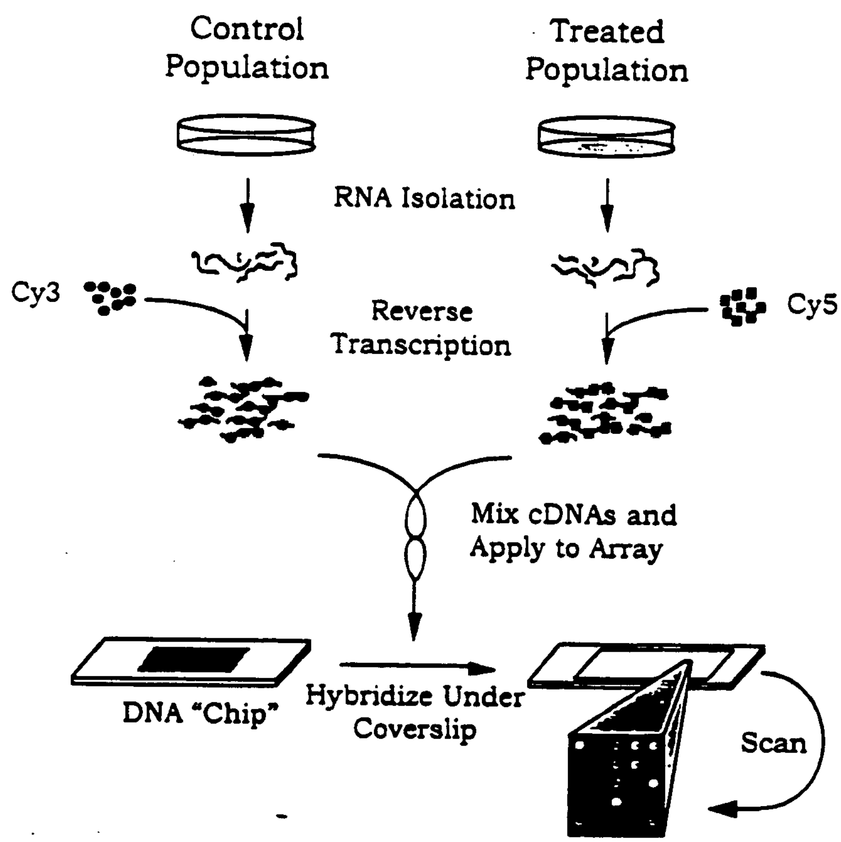


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

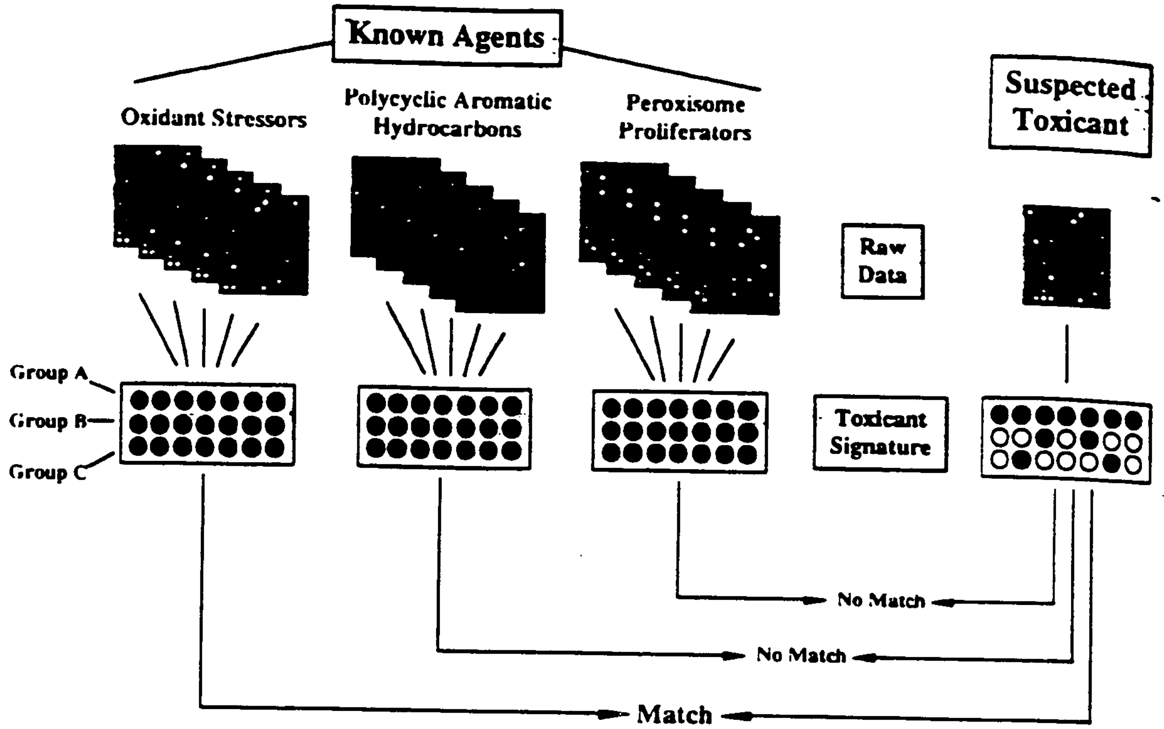


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabadopsis* could also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health ~~have made~~ laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257 967–971
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270 484–487
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270 467–470
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996;93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref. No. 1

XENOBIOTICA, 1999, VOL. 29, NO. 7, 655–691

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡ and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey, Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

UK.

ates network to the online edition.

nose University Press The online

ammes and details on advertising

venue, Emont, New York 11003
New York 11003

he pound sterling price applies. All
Canada, Mexico, India, Japan and
be made by sterling cheque, dollar

Post Bag No 8, Saket, New Delhi

PA 19106
8PR

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Heyningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

Xenobiotica ISSN 0049-8254 print/ISSN 1366-5928 online © 1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/xen.htm http://www.taylorandfrancis.com/JNLS/xen.htm

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate

ation of specific and
of cancer or previous
:ers.
enzymes (including
ucible by drugs and
lving transcriptional
it additional cellular
on. Accordingly, the
omplement of genes
the development of
of enzyme induction
of chemical-induced
adverse reactions to
n, some of which are
ical phenomenon *per*
profiling technologies
pools in target tissues
nisms of xenobiotic-
on in target tissues is
n generated in the
identification of toxic
ess and contributing
le, if the gene profile
rized *in vivo* could be
ative of all new drug
of toxicity, thereby
on of such toxicants.
ctionality of all genes
ger term goal, as the
ss their functionality
ds a *pattern* of gene
tched to that of well-
e *in vivo* similarities
a platform for more
beginning to gain
mercially producing
oxicity assessment of
es, some of which are
tically-related to any
l in broad-spectrum
gene arrays are now
nges in growth factor
: chemically-induced

changes presents a
s of development and
empting this difficult
methods have been
cts that demonstrate

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

(1) Differential screening,
(2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
(3) Differential display (DD),
(4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
(5) Gene expression arrays, and
(6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Daguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

)NA sequences from
:pared from normal.
ie filter replicas are
ır otherwise) labelled
mRNA populations.
d cell population will
'om the treated cells.
:an be used to probe
; which are only up-
)avis (1979) screened
bes in order to obtain
groundbreaking in its
nsuming, as up to 2
ıich are differentially
venient way to check
en completed.

d the success of early
9) soon gave rise to a
: to be developed was
:e below). In general,
ne population (tester)
by separation of the
: hybridized common
and through the use

: physical separation
:ıes. Several methods
tite chromatography
: and Dinauer 1990)
approach, common
\ (from control cells)
ography, as hydroxy-
ınabsorbed cDNA is
ferentially expressed
irectly as a probe to
.. 1984, Hedrick *et al.*
re 1.
sensitivity enhancing
me of the problems
Daguid and Dinauer
ry systems as a means
both the control and
daptor ('oligovector',

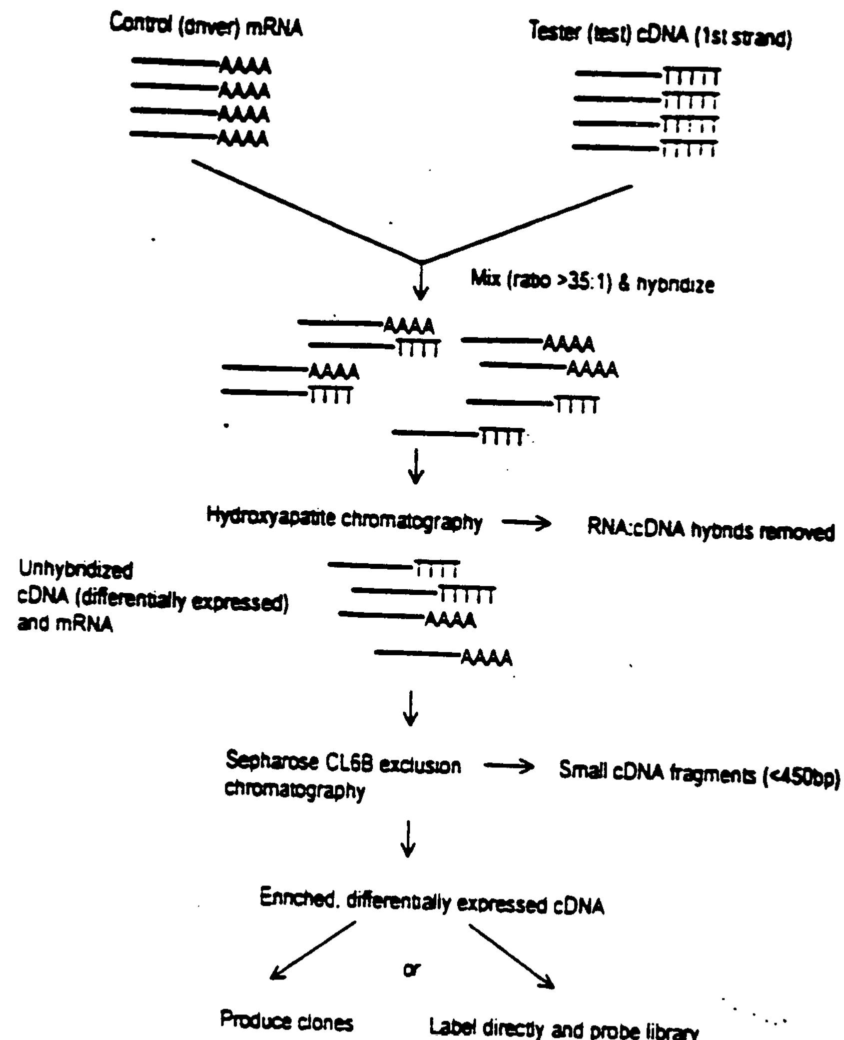


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/altered (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/altered population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligovector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

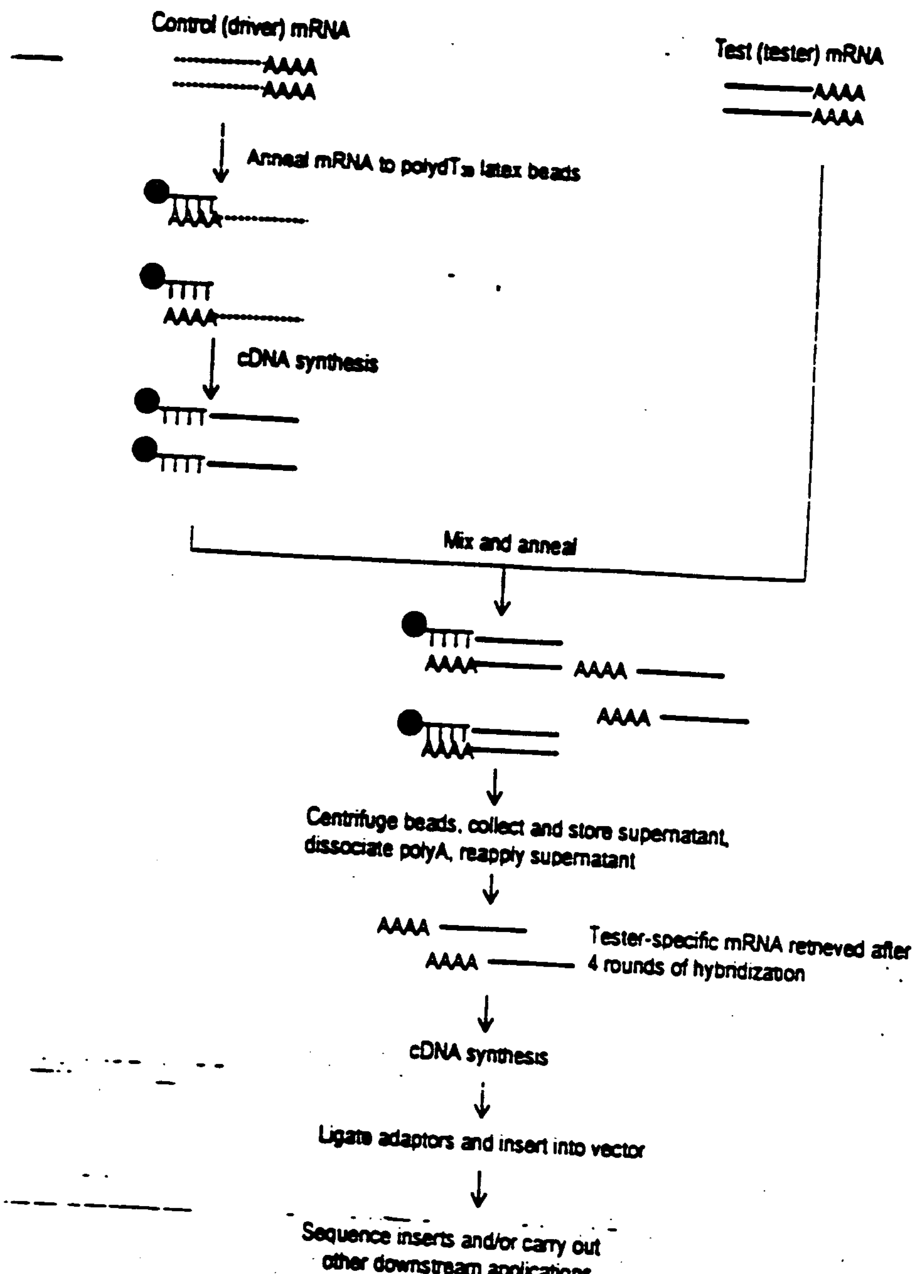


Figure 2. The use of oligodT$_{30}$ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/altered (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara *et al.* (1991).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT_{30}) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT_{30} forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT_{30} population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 μg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promotor sequence

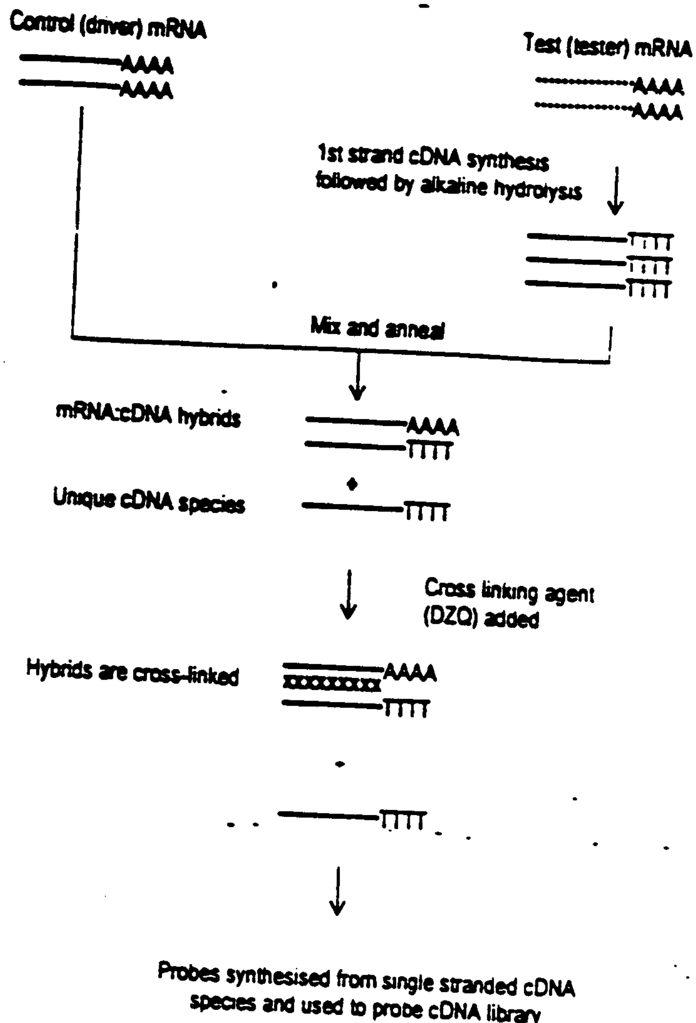


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diaziridinyl-1,4-benzoquinone (DZQ) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/μg total RNA
Abundant	12000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11000	0.004	0.002

— Modified from Bertioli *et al.* (1995).

at the 5′ end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5′ adaptors at each end of the DNA duplex and can, thus, be filled in at both 3′ ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Shatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

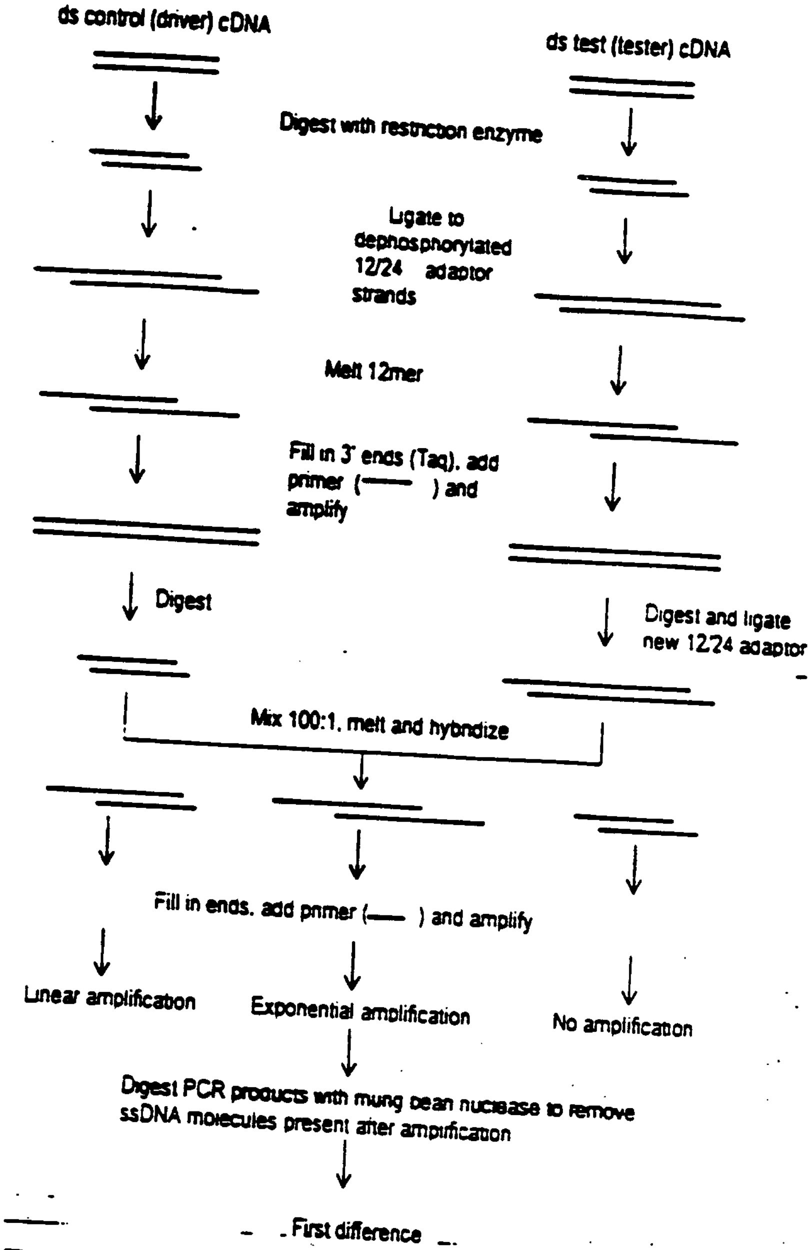


Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *Dpn*II. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3'ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *Dpn*II. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3' ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added ~~to two~~ portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

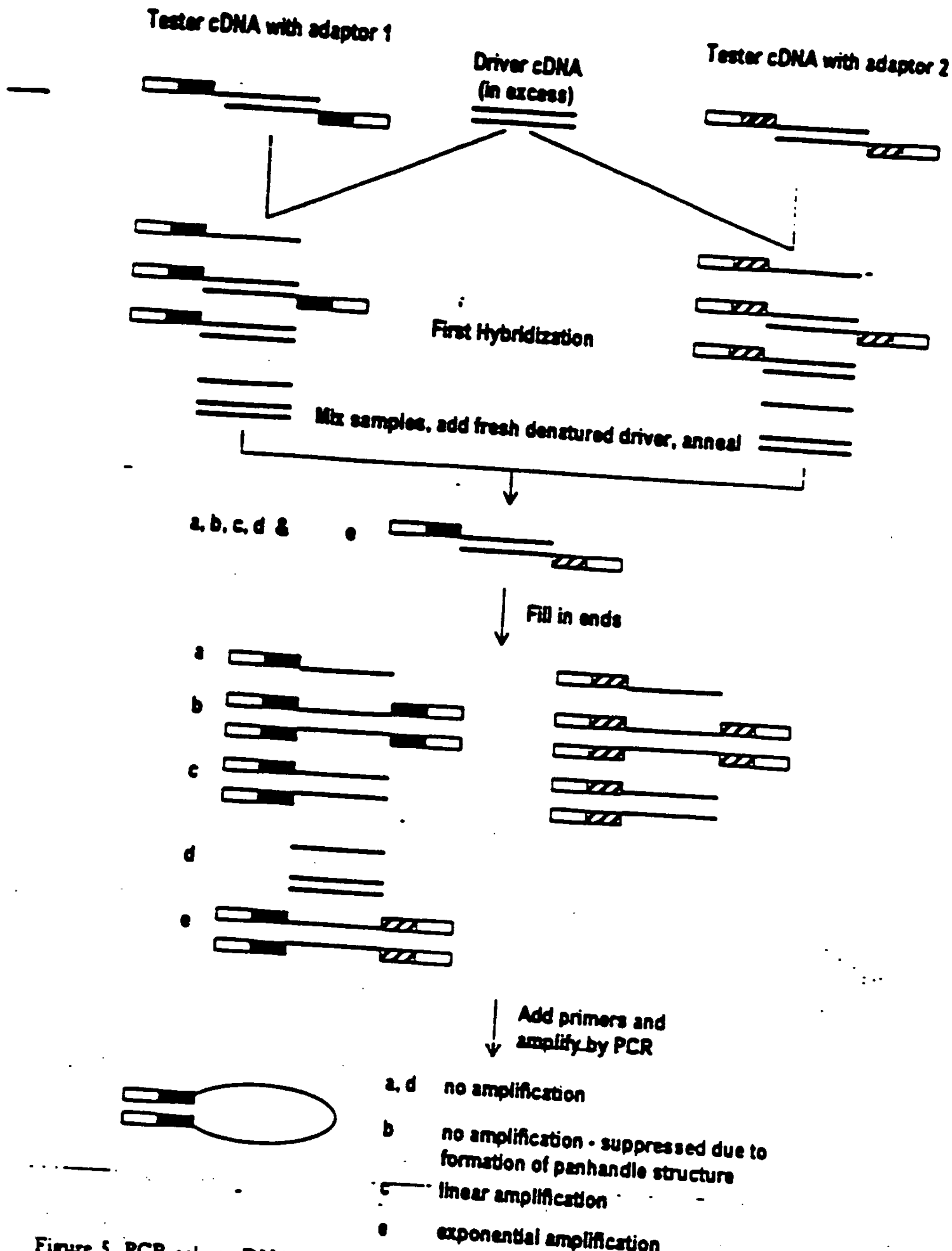
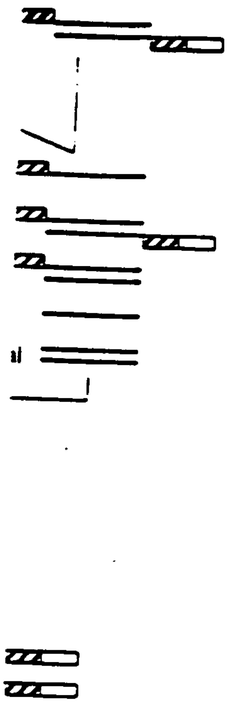


Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996), with permission.

r cDNA with adaptor 2



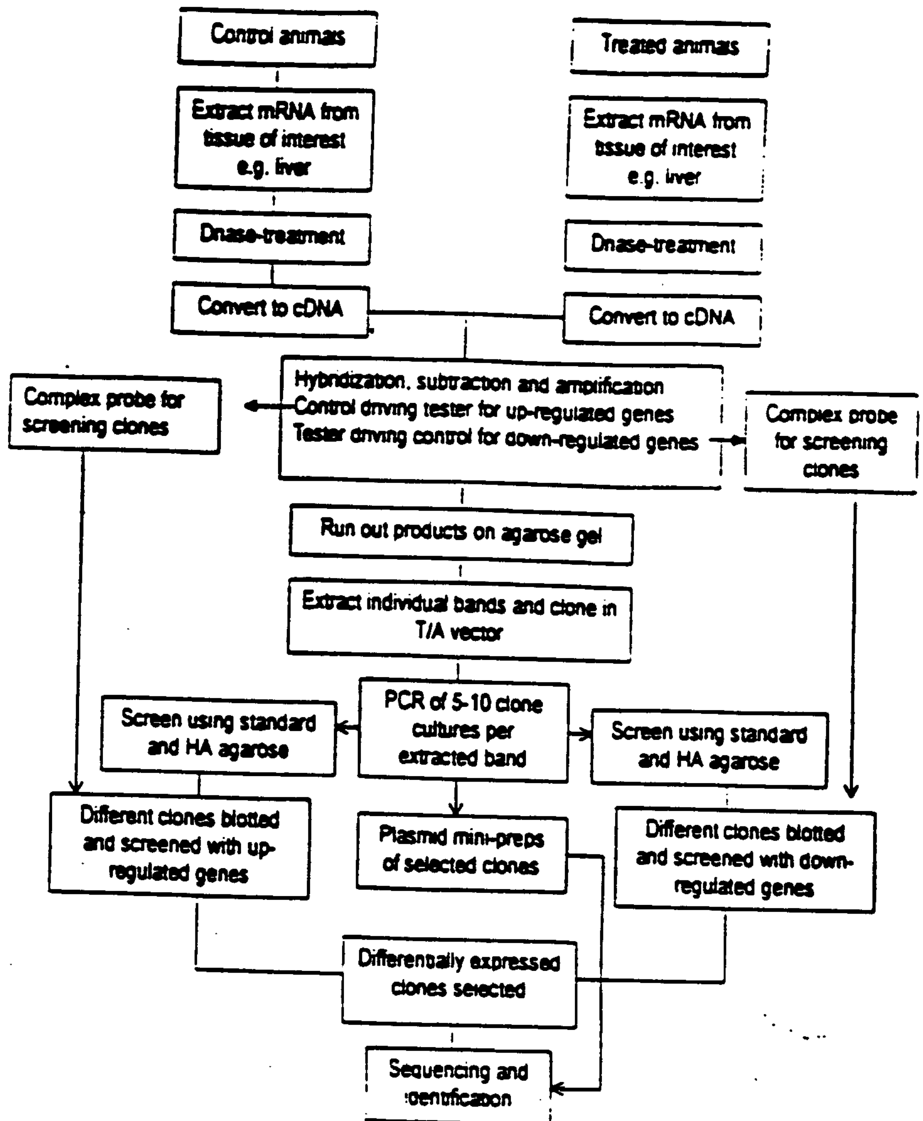


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

due to
ure

excess of driver cDNA is
ed and allowed to hybridize
d abundant molecules; and
not differentially expressed
dization, the two primary
ed driver can also be added
quences. Type e molecules
plified using two rounds of
d directly or cloned into a
et al. (1996) and Gurskaya

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

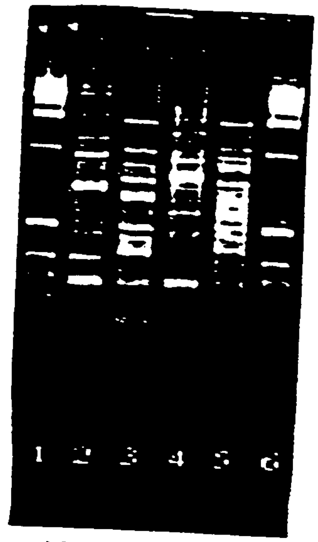


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy,14–643 treatment; 3—genes downregulated following Wy,14–643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al*. (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy,14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al*. 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete funtional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

:atment with WY-14,643 or
was used to generate the
(Clontech). Lane: 1—1kb
ıes downregulated following
rbital treatment; 3—genes
Reproduced from Rockett et

ıbtained. For example,
ıogen Wy,14,643, up-
in the rat (a sensitive
·s in the guinea pig, a
blished observations).
down-regulated in the
ely named TAPA-1) is
ırge number of cellular
:ifferentiation (Levy et
:ent in the phenomena
ıt is intriguing, and
:ifferentially regulated
.ae of this approach is
atabase sequences, but
genes of completely
rall assessment of the
g the lack of complete
gene profiling studies
ı xenobiotic challenge,
for further detailed

y primed PCR' (Liang
rred to as 'differential

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 *H. sapiens* (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobulin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNa

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	*M.musculus* C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	*M. musculus* Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT *M. musculus* (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML *M. musculus* (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 *M. musculus* (EST)
10 (550)	73.8%	Soares mouse NML *M. musculus* (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 *H. sapiens* (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 *M. musculus* (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	*R. norvegicus* RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5' $(dT_{11})CA$ 3' (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

(1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).

(2) The cDNAs obtained often only represent the extreme 3' end of the mRNA (often the 3'-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3' end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.

(3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather then total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

d may be recovered for
peed with which it can
:k to make and identify

ethods of priming the
with a 2-base 'anchor'
92). Alternatively, an
is (Welsh *et al.* 1992).
AP' (RNA Arbitrarily
PCR products may be
frames. In addition, it
nany bacterial mRNAs
erse transcription and
:th an arbitrary primer
compared to *random*
osition). The resulting
on the system (primer
ually includes 50–100
mbination of different
species from a cell can
pulations are analysed
n can be identified and
lysis.
l today for identifying
ceived disadvantages:

mRNAs (Bertioli *et al.*
id the isolation of very
stances (Guimeraes *et*

3' end of the mRNA
ot always be the case
uded in Genbank and
DD cannot always be
ed.
play often cannot be
in up to 70% of cases
reduce false positives,
and Denman 1997),
urse (Burn *et al.* 1994)
uced and two induced
ported that the use of
positives arising from

aknesses of the DD
: al. (1996) and from

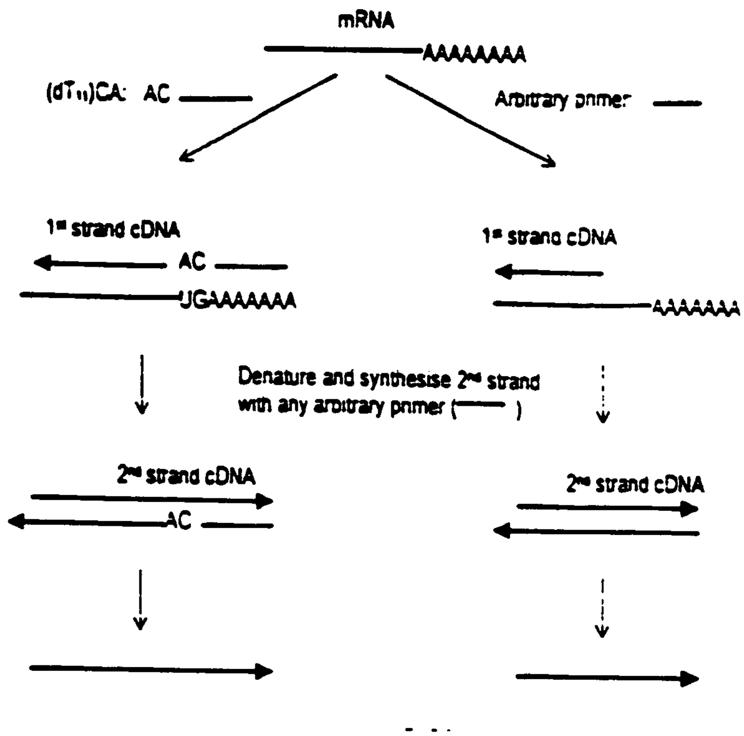


Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT$_{11}$NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatonation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

split into two and different adaptors ligated to the 5′ ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5′ digestion products. The use of restricted 3′-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3′ cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobolized 3′-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

ch group. Incorporated
·striction enzyme—one
recognition sequence.
·n with the IIS enzyme.
ıre released. The two
amplified products are
:atomers are formed in
: hundreds of gene tags
re, the number of times
·ement of that gene's
litates identification of

hnical difficulty of the
ased towards abundant
·nomic setting and has
date.

isolating differentially
·avsky (1995). In this
ligo(dT) primers. The
ease and captured with
· unwanted 5′ digestion
e the complexity of the
·ecies is represented by
to facilitate subsequent
out with one adaptor-
nplified population is
aline dissociation. The
ferent adaptor-specific
immobilized 3′ cDNA
·triction endonucleases
e result is a fingerprint
quential digests used).
› identify differentially
gel and cloned. The
reproducible, and the
involved in the final
an rarely resolve more
0 or more which are
se of 2-D gels such as
al. (1991) may help to

se fragments was later
instead of sequential
. these authors simply
·tions without further

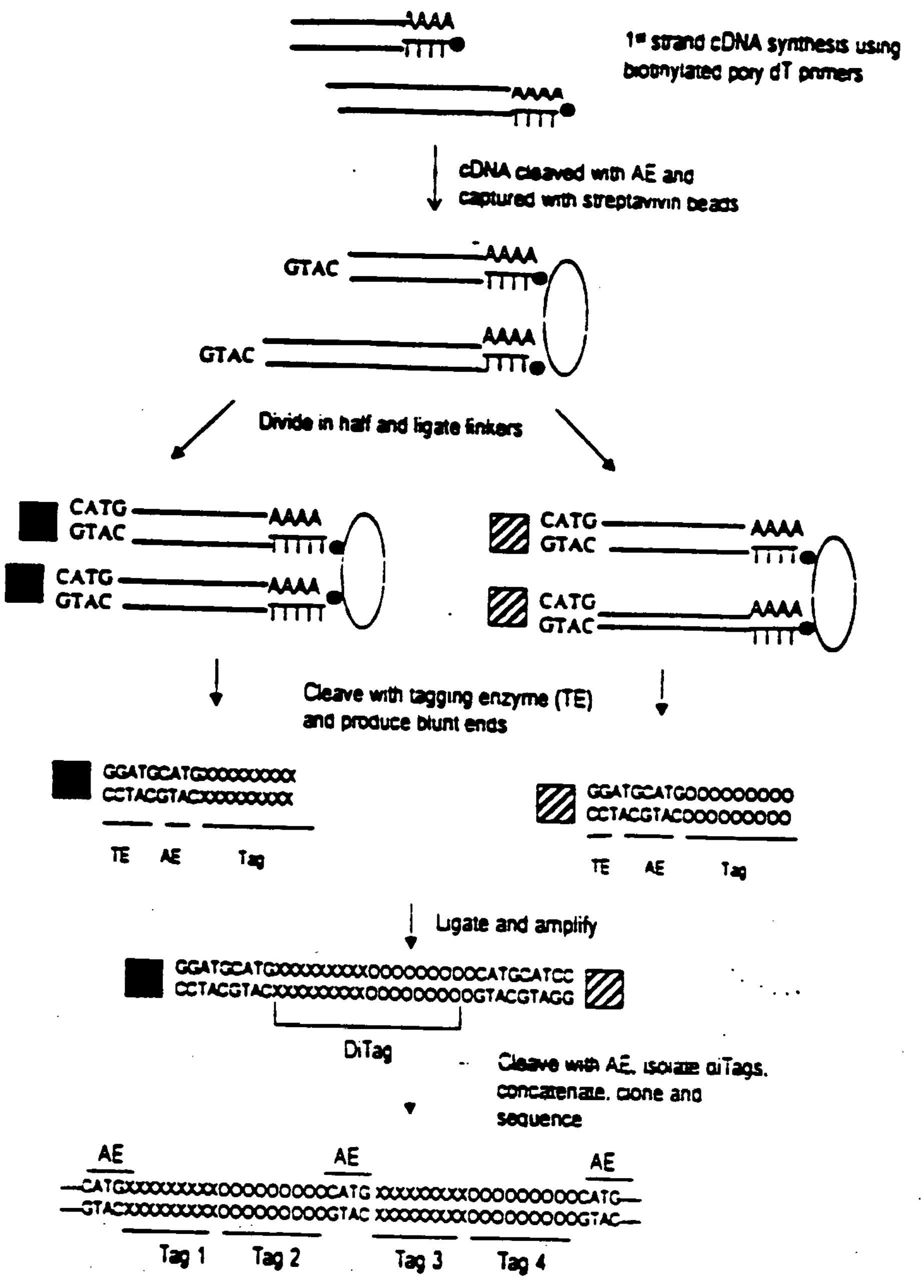


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3′ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the ditags isolated from the linkers using PAGE. The ditags are then ligated (during which process, concatenization occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half

.at it takes a great deal
:n that they are indeed
. tissue. Normally, the
PCR. Even so, each of
: goal of rapid analysis
by the development of
5, Schena *et al.* 1996),
ential gene expression
iss 'chips' containing
:tiple copies of part of
y proven involvement
her cellular processes.
ne and animal species.
and a few companies
tech Laboratories and
:ds or even thousands
cDNA from the test
When analysed with
quantitative means to
ions. Of course, there
hich are in the array
h to elucidating the
pment system may be
directly identify and
lations, and an open
:rentially expressed.
ber of gene fragments
ported gridding up to
se high density chip-
roduced off-the-shelf
pid determination of
:ts. Aside from their
:g and probing DNA
e newer micro-array
eproducible between
e resolved within the

:essed genes
:lones obtained from
il identity (putative
: a rapid and efficient
ate profiles of gene-
Adams *et al.* (1991),
imated that there are
epresenting over half

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatzis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at http://www.tigr.org) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatzis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdisection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: http://www.ncbi.nlm.nih.gov/ncicgap/intro.html). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,

e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromized tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005 % of the total population (table 1). Bertioli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA

. 1998, Kas-Deelen *et*
. Rogler *et al.* 1998).
ıis issue unimportant.
'ing altered expression
ıon. After all, since all
pes which intimately
each cell type could in
nolecular mechanisms
growth. It is perhaps
ments using *in vivo* as
ientical cells probably
nolecular changes that

.al biological variation
.dels are being used. It
erent ways to identical
ebrisoquine oxidation
5 and determines the
ıard 1993, Meyer and
d complex, but allelic
cal and mental health
nses. Careful thought
dy and to the possible
effect of this can be
ıd unimportant minor
vidual animals, thus
r mechanisms of the
ıs may be of utmost
ccumb to or resist the

a high percentage of

uggesting that mam-
pecies at any one time
), although figures as
Hedrick *et al.* (1984)
to the rare abundance
n table 1.
been compared with
ıt not all differentially
ticular, rare messages
not easily recovered
ıg, as the majority of
population (table 1).
lates (heterogeneous
able to detect mRNA

species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000 × smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon-γ-stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3′ restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3′ polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between

ally effective—proving
numbers of artificial
e rare messages already
· models will genuinely
s. In addition, there are
ple, mRNAs may have
amplification by PCR-
circumstances not all
evelopment, deadenyl-
d Steitz 1998), whilst
Hsp70 (and perhaps,
avalle *et al.* 1994). The
e efficiency of systems
cy of any system also
tial display techniques
› isolate mRNA that is
are used to prime first
ribed to some degree
It has been shown, at
can lead to inefficient
Subtraction kit user
o likewise in other SH
ction amplification step
me sequences amplify

the temporal factor. It
ily interrogate a cell at
genes showing altered
disease processes and
scades of signalling,
es which are switched
vital information may
information about the
gy can be derived for
cular interest to the
time point analysis is
hich, of course, adds

ssue of how large the
gene in question with
the isolation of genes
reported using SSH
nstrating a change in
here is a 'grey zone'—
of isolation between

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hessle, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

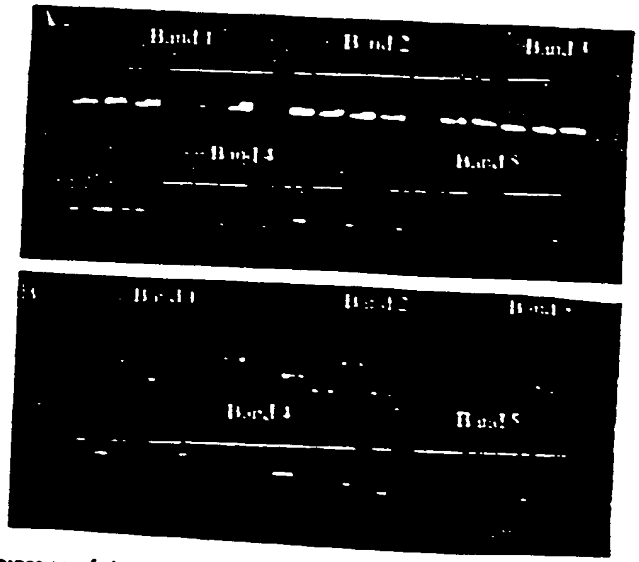


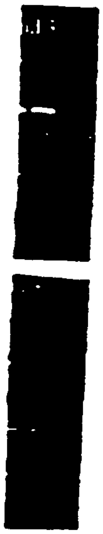
Figure 10 Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).



IA-red. Bands of decreasing
ı subtractive hybridization
each cloned band and their
high resolution 2 % agarose
·ed. With few exceptions, all
·er. the presence of HA-red
he percentage of GC within
:ies within each band. For
ıe the same size, at least four

rd gel should indicate
ld separate otherwise
ase content. Geisinger
entifying DD-derived
:s laboratory on clones

he differential display
:ied out in a standard
:ted and incorporated

there being different
AT content. However,
-again, one might use
)GGE) or temperature
he contents of a band,
or on the reamplified

ıes to visualize large
oblem in that, in terms
ands. One approach to
bed by Uitterlinden *et*

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30 %) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to overstain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 bp on the same gel, which leads to compromize in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to included certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitemate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 or its as yet undiscovered sister X_2? For example, using SSH, part of a gene was isolated,

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

(1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
(2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using mulitchannel pipettes, 96+-well plates and

appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN-γ, Frye *et al.* 1989), β-actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β-2-microglobulin (β-2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al.* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using the differential display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as decribed extensively herein, but rather protein–protein, protein–DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

cannot hope to answer
.entification. regulatory
ıolecular mechanism of
regulation and function
tract sense. differential
s of a fixed moment in
attle and the placement
but is asked to try and
:t did from a few still
ıe battle, the Historian
and their commanding
ł. He must examine the
the prevailing weather
'o be forthcoming, the
her techniques, such as
, mutation analysis and
' has emphasized the
sidered in isolation and
·d in combination with
n gels from isoelectric
D-maps using capillary
ention as many of the
ɔlve changes in mRNA
tein, protein–DNA and
ınctional genomics or

y, it is clear that many
ıracterizing the genetic
ɪpment and in response
.ta, such profiling will
ponse, and in the long
ıiomarkers for different
: potential medical and
anges are aimost im-
ındicate the family or
·sed to plus the length
ɔst prudent treatment.
ntical cancers, provide
n, perhaps indicate the

ıe next century and the
ɔntinuing development
will ensure that this
disease processes.

Sally Damey and Chris
: manuscript prior to
:e with the policy of the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. and Ventor, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

An, G., Luo, G., Veltri, R. W. and O'Hara, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, 20, 342–346.

Axel, R., Feigelson, P. and Schultz, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, 7, 247–254.

Band, V. and Sager, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249–1253.

Bauer, D., Muller, H., Reich, J., Riedel, H., Ahrenkiel, V., Warthoe, P. and Strauss, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, 21, 4272–4280.

Bertioli, D. J., Schlichter, U. H. A., Adams, M. J., Burrows, P. R., Steinbiss, H.-H. and Antoniw, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520–4523.

Bravo, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, 1, 37–46.

Burn, T. C., Petrovick, M. S., Hohaus, S., Rollins, B. J. and Tenen, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776–2783.

Cao, J., Cai, X., Zheng, L., Geng, L., Shi, Z., Pao, C. C. and Zheng, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447–451.

Cassidy, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13–20.

Chang, G. W. and Terzaghi-Howe, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445–4452.

Chen, J., Schwartz, D. A., Young, T. A., Norris, J. S. and Yager, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, 17, 2783–2786.

Chen, J. J. W. and Peck, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, 24, 793–794.

ClonTechniques, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18–19.

ClonTechniques, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15–16.

Davis, M. M., Cohen, D. I., Nielsen, E. A., Steinmetz, M., Paul, W. E. and Hood, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, 81, 2194–2198.

Dellavalle, R. P., Peterson, R. and Lindquist, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in Drosophila melanogaster. *Molecular and Cell Biology*, 14, 3646–3659.

DeRisi, J. L., Vashwanath, R. L. and Brown, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Diatchenko, L., Lau, Y.-F. C., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, 93, 6025–6030.

Dogra, S. C., Whitelaw, M. L. and May, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1–9.

Duguid, J. R. and Dinauer, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789–2792.

Dunbar, P. R., Ogg, G. S., Chen, J., Rust, N., van der Bruggen, P. and Cerundolo, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, 26, 413–416.

FITZPATRICK, D. R., GERMAIN-LEE, E. and VALLE, D., 1995. Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, 27, 457–466.

FOSS, D. L., BAARSCH, M. J. and MURTAUGH, M. P., 1998. Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, 9, 67–78.

FRYE, R. A., BENZ, C. C. and LIU, E., 1989. Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, 4, 1153–1157.

GEISINGER, A., RODRIGUEZ, R., ROMERO, V. and WETTSTEIN R., 1997. A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, http://tto.trends.com, document T01110.

GRESS, T. M., HOHEISEL, J. D., LENNON, G. G., ZEHETNER, G. and LEHRACH, H., 1992. Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, 3, 609–619.

GRIFFIN, G. and KRISHNA, S., 1998. Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, 32, 195–198.

GROENINK, M. and LEEGWATER, A. C. J., 1996. Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, XI, 23–24.

GUIMARAES, M. J., BAZAN, J. F., ZLOTNIK, A., WILES, M. V., GRIMALDI, J. C., LEE, F. and MCCLANAHAN, T., 1995b. A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, 121, 3335–3346.

GUIMERAES, M. J., LEE, F., ZLOTNIK, A. and MCCLANAHAN, T., 1995a. Differential display by PCR: novel findings and applications. *Nucleic Acids Research*, 23, 1832–1833.

GURSKAYA, N. G., DIATCHENKO, L., CHENCHIK, P. D., SIEBERT, P. D., KHASPEKOV, G. L., LUKYANOV, K. A., VAGNER, L. L., ERMOLAEVA, O. D., LUKYANOV, S. A. and SVERDLOV, E. D., 1996. Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-Myrystate 13-Acetate. *Analytical Biochemistry*, 240, 90–97.

HAMPSON, I. N. and HAMPSON, L., 1997. CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), 23, 22–24.

HAMPSON, I. N., HAMPSON, L. and DEXTER, T. M., 1996. Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*, 24, 4832–4835.

HAMPSON, I. N., POPE, L., COWLING, G. J. and DEXTER, T. M., 1992. Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, 20, 2899.

HARA, E., KATO, T., NAKADA, S., SEKIYA, S. and ODA, K., 1991. Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, 19, 7097–7104.

HATADA, I., HAYASHIZAKI, Y., HIROTSUNE, S., KOMATSUBARA, H. and MUKAI, T., 1991. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences (USA)*, 88, 9523–9527.

HECHT, N., 1998. Molecular mechanisms of male sperm cell differentiation. *Bioessays*, 20, 555–561.

HEDRICK, S., COHEN, D. I., NIELSEN, E. A. and DAVIS, M. E., 1984. Isolation of T cell-specific membrane-associated proteins. *Nature*, 308, 149–153.

HERTZ, R., SECKBACH, M., ZAKIN, M. M. and BAR-TANA, J., 1996. Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, 271, 218–224.

HELVAL, J. P. V., CLARK, G. C., KOHN, M. C., TRITSCHER, A. M., GREENLEE, W. F., LUCIER, G. W. and BELL, D. A., 1994. Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transciptase-polymerase chain reaction. *Cancer Research*, 54, 62–68.

HILLIER, L. D., LENNON, G., BECKER, M., BONALDO, M. F., CHIAPELLI, B., CHISSOE, S., DIETRICH, N., DUBUQUE, T., FAVELLO, A., GISH, W., HAWKINS, M., HULTMAN, M., KUCABA, T., LACY, M., LE, M., LE, N., MARDIS, E., MOORE, B., MORRIS, M., PARSONS, J., PRANGE, C., RIFKIN, L., ROHLFING, T., SCHELLENBERG, K., SOARES, M. B., TAN, F., THIERRY-MIEG, J., TREVASKIS, E., UNDERWOOD, K., WOHLDMAN, P., WATERSTON, R., WILSON, R and MARRA, M., 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, 6, 807–828.

HUBANK, M. and SCHATZ, D. G., 1994. Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, 22, 5640–5648.

HUNTER, T., 1991. Cooperation between oncogenes. *Cell*, 64, 249–270.

IVANOVA, N. B. and BELYAVSKY, A. V., 1995. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, 23, 2954–2958.

JAMES, B. D. and HIGGINS, S. J, 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).

KAS-DEELEN, A. M., HARMSEN, M. C., DE MAAR, E. F. and VAN SON, W. J, 1998. A sensitive method for

nd characterisation of rat and
-oA hydratase. *Genomics*, 27,

n of hypoxanthine phospho-
d beta-actin mRNA expression
78.

ones by differential polymerase

simple method for screening
inds. *Elsevier Trends Journals*

RACH, H., 1992. Hybridisation
is derived from whole tissues.

Journal of the Royal College of

ily genes associated with liver
iotechniques, XI, 23–24.
IMALDI, J. C., LEE, F. and
poietic development in the yolk

993a. Differential display by
1832–1833.
CHASPEKOV, G. L., LUKYANOV,
and SVERDLOV, E. D., 1996.
of polymerase chain reaction:
and phorbol 12-Myrystate 13-

cloning made easy. *Life Science*

andom oligonucleotide primed
active cDNA cloning. *Nucleic*

mical cross linking subtraction
disation probes. *Nucleic Acids*

tractive cDNA cloning using
c to undifferentiated human

MUKAI, T., 1991. A genomic
landmarks. *Proceedings of the*

on. *Bioessays*, 20, 555–561.
Isolation of T cell-specific

scriptional suppression of the
l of Biological Chemistry, 271.

EE, W. F., LUCIER, G. W. and
e-response relationships using
er Research, 54, 62–68.
B., CHISSOE, S., DIETRICH, N.,
L., KUCABA, T., LACY, M., LE,
ICE, C., RIFKIN, L., ROHLFING,
TREVASKIS, E., UNDERWOOD,
1996. Generation and analysis
07–528
expression by representational

expressed genes by restriction
Research, 23, 2954–2958.
d: IRL Press Ltd).
. 1998. A sensitive method for

quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, **5**, 622–626.

KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), 10, 50–51.

KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, 7, 1611–1618.

KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, **18**, 5705–5711.

LAKE, B. G., EVANS, J. G., CUNNINGHAME, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, **101**, 241–248.

LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guiea pig and marmoset. *Toxicology and Applied Pharmacology*, 99, 148–160.

LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, 9, 60–77.

LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, **16**, 89–109.

LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.

LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, **52**, 6966–6968.

LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, **21**, 3269–3275.

LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOE, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, **254**, 304–321.

LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, **23**, 3244–3251.

LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, **259**, 946–951.

LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, **18**, 200–202.

LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Techology*, **17**, 84–94.

MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepaic cytochrome P4504A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, **7**, 183–191.

MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, **53**, 39–51.

MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, 24, 1504–1507.

MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), 10.19–20.

MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, 11, 242–246.

MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, **88**, 29–36.

MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269–296.

MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, **28**, 437–447.

MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the mdr-1 gene expression. *Biochemistry*, **29**, 10351–10356.

NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYEREISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, **6**, 1–42.

NISHIO, Y., AIELLO, L. P. and KING, G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, 8, 103–106.

O'NEILL, M. J. and SINCLAIR, A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, 25, 2681–2682.

ORTON, T. C., ADAM, H. K., BENTLEY, M., HOLLOWAY, B. and TUCKER, M. J., 1984, Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, 73, 138–151.

PELKONEN, O., MAENPAA, J., TAAVITSAINEN, P., RAUTIO, A. and RAUNIO, H., 1998, Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica*, 28, 1203–1253.

PHILIPS, S. M., BENDALL, A. J. and RAMSHAW, I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, 82, 199–203.

PRASHAR, Y. and WEISSMAN, S. M., 1996, Analysis of differential gene expression by display of 3'end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences (USA)*, 93, 659–663.

RAGNO, S., ESTRADA, I., BUTLER, R. and COLSTON, M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, 57, 143–146.

RAMANA, K. V. and KOHLI, K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, 36, 437–446.

RICHARD, L., VELASCO, P. and DETMAR, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, 240, 1–6.

ROCKETT, J. C., ESDAILE, D. J. and GIBSON, G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug Metabolism and Pharmacokinetics*, 22, 329–333.

RODRICKS, J. V. and TURNBULL, D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, 3, 197–212.

ROGLER, G., HAUSMANN, M., VOGL, D., ASCHENBRENNER, E., ANDUS, T., FALK, W., ANDREESEN, R., SCHOLMERICH, J. and GROSS, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, 112, 205–215.

ROHN, W. M., LEE, Y. J. and BENVENISTE, E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, 16, 311–330.

RUDIN, C. M. and THOMPSON, C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, 25, 435–446.

SAKAGUCHI, N., BERGER, C. N. and MELCHERS, F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, 5, 2139–2147.

SAMBROOK, J., FRITSCH, E. F. and MANIATIS, T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Fergusen (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6–37.

SARGENT, T. D. and DAWID, I. B., 1983, Differential gene expression in the gastrula of Xenopus laevis. *Science*, 222, 135–139.

SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O. and DAVIS, R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences (USA)*, 93, 10614–10619.

SCHNEIDER, C., KING, R. M. and PHILIPSON, L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, 54, 787–793.

SCHNEIDER-MAUNOURY, S., GILARDI-HEBENSTREIT, P. and CHARNAY, P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, 321, 819–834.

SEMENZA, G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, 3, 180–199.

SEWALL, C. H., BELL, D. A., CLARK, G. C., TRITSCHER, A. M., TULLY, D. B., VANDEN HEUVEL, J. and LUCIER, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, 41, 1829–1834.

SINGH, N., AGRAWAL, S. and RASTOGI, A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, 3, 41–49.

SMITH, N. R., LI, A., ALDERSLEY, M., HIGH, A. S., MARKHAM, A. F. and ROBINSON, P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids* Research, 25, 3552–3554.

SOMPAYRAC, L., JANE, S., BURN, T. C., TENEN, D. G. and DANNA, K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, 23, 4738–4739.

ST JOHN, T. P. and DAVIS, R. W., 1979, Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridisation. *Cell*, 16, 443–452.

SUN, Y., HEGAMYER, G. and COLBURN, N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, 54, 1139–1144.

s in bovine aortic smooth muscle
8. 103–106.
s by representational difference

M. J., 1984. Clobuzarit: species
of the liver following chronic
51.
NIO, H., 1998. Inhibition and
tica, 28, 1203–1253.
of genes associated with high
of the National Cancer Institute.

expression by display of 3'end
cademy of Sciences (USA), 93.

egulation of macrophage gene
munology Letters, 57, 143–146.
me P450—an overview. Indian

gnetic protocol for the selective
endothelial cells. Experimental

ar profiling of non-genotoxic
tion-polymerase chain reaction
acokinetics, 22, 329–333.
peroxisomes and peroxisome

T., Falk, W., Andreesen, R.,
pic characterization of colonic
5.
ss II MHC expression. Critical

turation. Seminars in Oncology.

DNA copy of an RNA species

resis of DNA. In N. Ford, M.
anual, 2nd edition (New York:

the gastrula of Xenopus laevis.

s, R. W., 1996. Parallel human
000 genes Proceedings of the

expressed at growth arrest of

995. How to build a vertebrate
21, 81?–534
nanisms and pathophysiology.

D. B., Vanden Heuvel, J. and
ons for biomarkers. Clinical

d immunity: special reference
3, 41–49.
Robinson, P. A., 1997, Rapid
DDRT-PCR polyacrylamide

1995. Overcoming limitations
ch, 23, 4738–4739.
ucible DNA sequences from
n. Cell, 16, 443–452.
ing of five messenger RNAs
ouse epidermal cells: one is
ancer Research, 54, 1139–1144.

Sung, Y. J. and Denman, R. B., 1997. Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, 23, 462–464.

Sutton, G., White, O., Adams, M. and Kerlavage, A., 1995. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1, 9–19.

Suzuki, Y., Sekiya, T. and Hayashi, K., 1991. Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, 192, 82–84.

Syed, V., Gu, W. and Hecht, N. B., 1997. Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, 18, 264–273.

Uitterlinden, A. G., Slagboom, P., Knook, D. L. and Vijg, J., 1989. Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, 86, 2742–2746.

Ullman, K. S., Northrop, J. P., Verweij, C. L. and Crabtree, G. R., 1990. Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, 8, 421–452.

Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B. and Pastan, I., 1998. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, 95, 300–304.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W., 1995. Serial analysis of gene expression. *Science*, 270, 484–487.

Voeltz, G. K. and Steitz, J. A., 1998. AUUUA sequences direct mRNA deadenylation uncoupled from decay during Xenopus early development. *Molecular and Cell Biology*, 18, 7537–7545.

Vogelstein, B. and Kinzler, K. W., 1993. The multistep nature of cancer. *Trends in Genetics*, 9, 138–141.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997. A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13–14.

Wan, J. S., Sharp, S. J., Poirier, G. M.-C., Wagaman, P. C., Chambers, J., Pyati, J., Hom, Y.-L., Galindo, J. E., Huvar, A., Peterson, P. A., Jackson, M. R. and Erlander, M. G., 1996. Cloning differentially expressed mRNAs. *Nature Biotechnology*, 14, 1685–1691.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997. A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, 21, 13–14.

Wang, Z. and Brown, D. D. 1991. A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, 88, 11505–11509.

Wawer, C., Ruggeberg, H., Meyer, G. and Muyzer, G., 1995. A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, 23, 4928–4929.

Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D. and McClelland, M., 1992. Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, 20, 4965–4970.

Wong, H., Anderson, W. D., Cheng, T. and Riabowol, K. T., 1994. Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, 223, 251–258.

Wong, K. K. and McClelland, M., 1994. Stress-inducible gene of Salmonella typhimurium identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, 91, 639–643.

Wynford-Thomas, D., 1991. Oncogenes and anti-oncogenes: the molecular basis of tumour behaviour. *Journal of Pathology*, 165, 187–201.

Xhu, D., Chan, W. L., Leung, B. P., Huang, F. P., Wheeler, R., Piedrafita, D., Robinson, J. H. and Liew, F. Y., 1998. Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, 187, 787–794.

Yang, M. and Sytowski, A. J., 1996. Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, 237, 109–114.

Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. and Sakaki, Y., 1995. High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, 156, 207–213.

Zhao, X. J., Newsome, J. T. and Cihlar, R. L., 1998. Up-regulation of two *candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, 25, 121–129.

Zimmermann, C. R., Orr, W. C., Leclerc, R. F., Barnard, C. and Timberlake, W. E., 1980. Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, 21, 709–715.

ELSEVIER

Toxicology Letters 112-113 (2000) 467-471

Toxicology Letters

www.elsevier.com/locate/toxlet

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

0378-4274/00/$ - see front matter © 2000 Elsevier Science Ireland Ltd. All rights reserved.
PII: S0378-4274(99)00236-2

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

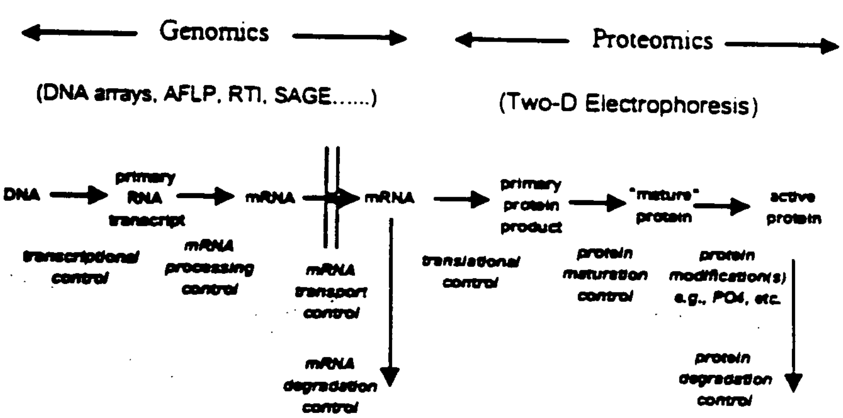


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

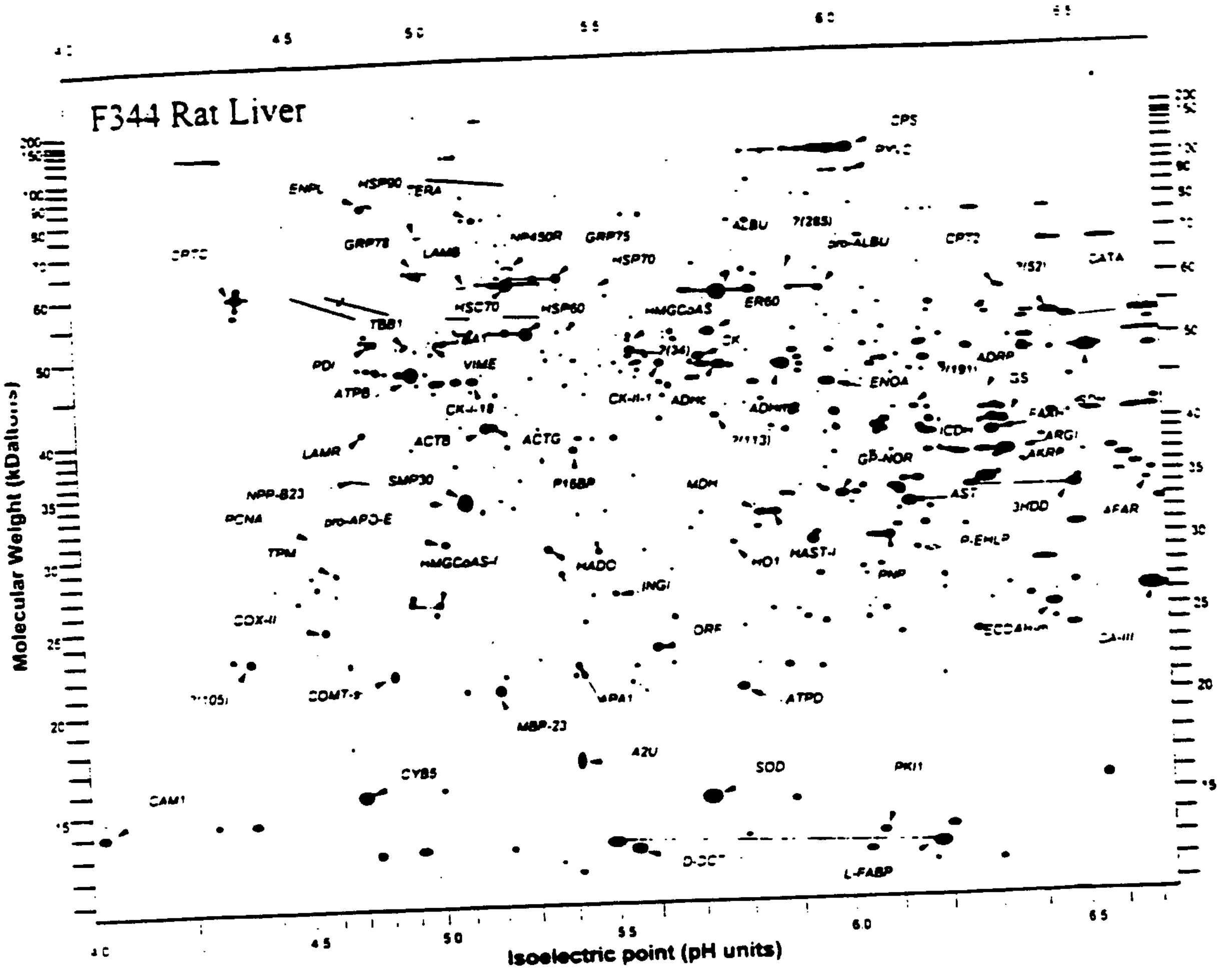


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug development, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raymackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref. No. 4

Subject: RE: [Fwd: Toxicology Chip]
Date: Mon. 3 Jul 2000 08:09:45 -0400
From: "Afshari.Cynthia" <afshari@niehs.nih.gov>
To: "'Diana Hamlet-Cox'" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at
http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox
response and basic cellular processes and added a set of clones and ESTs to
this. We have included a set of control genes (80+) that were selected by
the NHGRI because they did not change across a large set of array
experiments. However, we have found that some of these genes change
signficantly after tox treatments and are in the process of looking at the
variation of each of these 80+ genes across our experiments.
Our chips are constantly changing and being updated and we hope that our
data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> *From:* *Diana Hamlet-Cox*
> *Sent:* *Monday, June 26, 2000 8:52 PM*
> *To:* *afshari@niehs.nih.gov*
> *Subject:* *[Fwd: Toxicology Chip]*
>
> *Dear Dr. Afshari,*
>
> *Since I have not yet had a response from Bill Grigg, perhaps he was not*
> *the right person to contact.*
>
> *Can you help me in this matter? I don't need to know the sequences,*
> *necessarily, but I would like very much to know what types of sequences*
> *are being used, e.g., GPCRs (more specific?), ion channels, etc.*
>
> *Diana Hamlet-Cox*
>
> *-------- Original Message --------*
> *Subject: Toxicology Chip*
> *Date: Mon, 19 Jun 2000 18:31:48 -0700*
> *From: Diana Hamlet-Cox <dianahc@incyte.com>*
> *Organization: Incyte Pharmaceuticals*
> *To: grigg@niehs.nih.gov*
>
> *Dear Colleague:*
>
> *I am doing literature research on the use of expressed genes as*
> *pharmacotoxicology markers, and found the Press Release dated February*
> *29, 2000 regarding the work of the NIEHS in this area. I would like to*
> *know if there is a resource I can access (or you could provide?) that*
> *would give me a list of the 12,000 genes that are on your Human ToxChip*
> *Microarray. In particular, I am interested in the criteria used to*
> *select sequences for the ToxChip, including any control sequences*
> *included in the microarray.*
>
> *Thank you for your assistance in this request.*
>
> *Diana Hamlet-Cox, Ph.D.*
> *Incyte Genomics, Inc.*
>
> --
>
> ========================

> This email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> ========================
>
>

Docket No.: PF-0232-1 DIV
USSN: 09/782,390
Ref. No. 5

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties $-12/-1$ (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

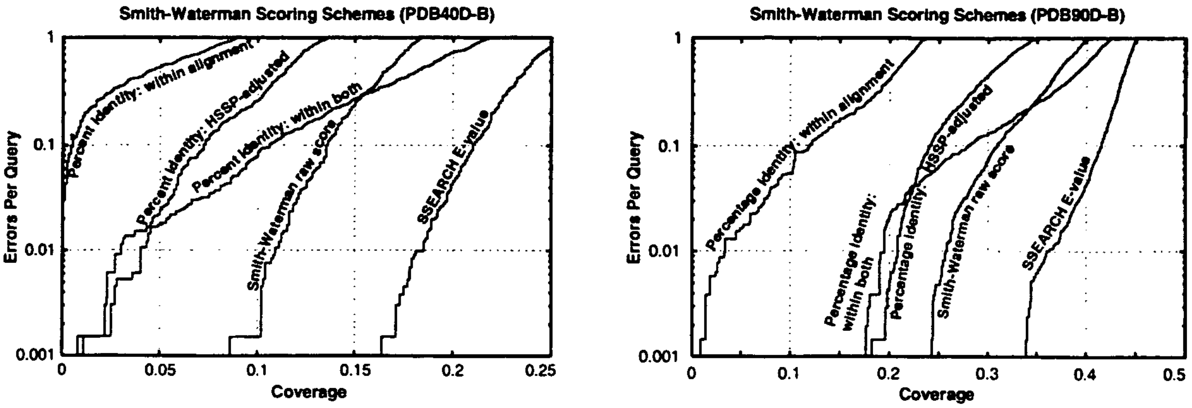


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

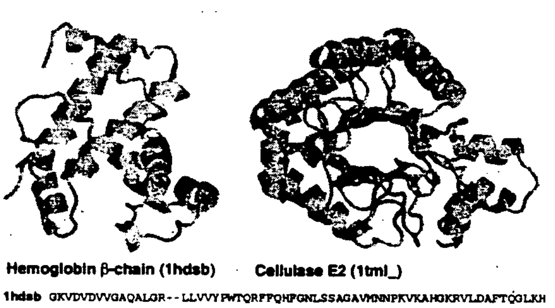


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

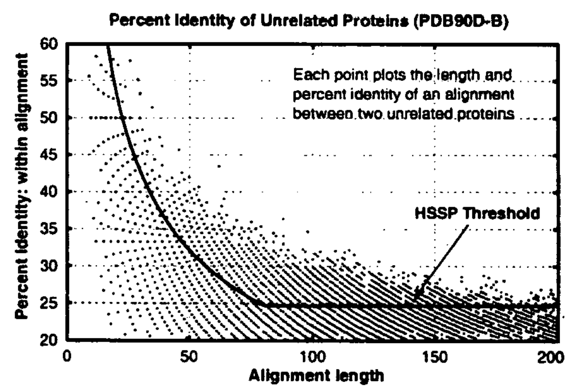


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

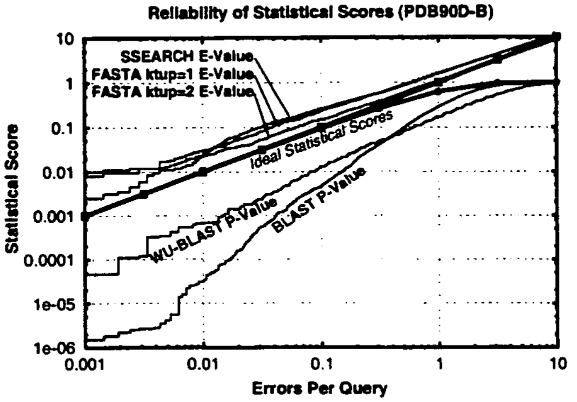


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

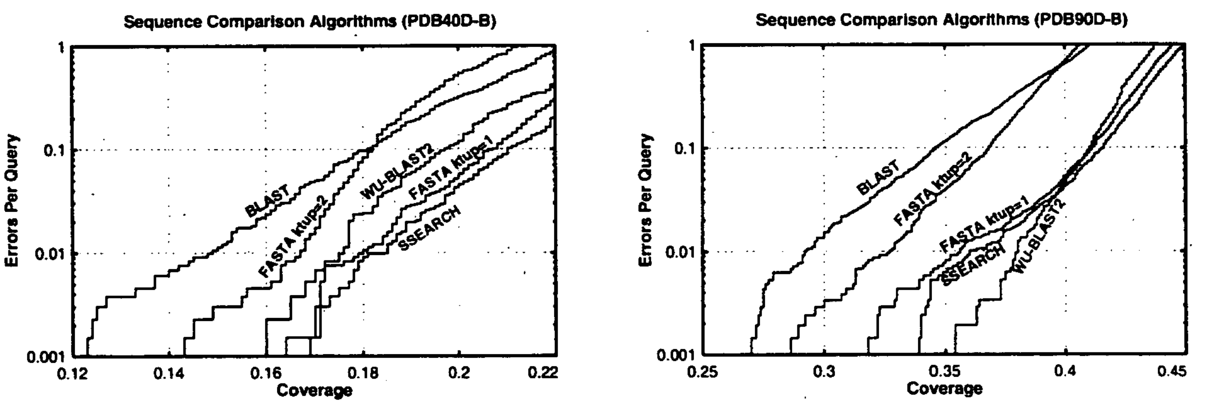


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

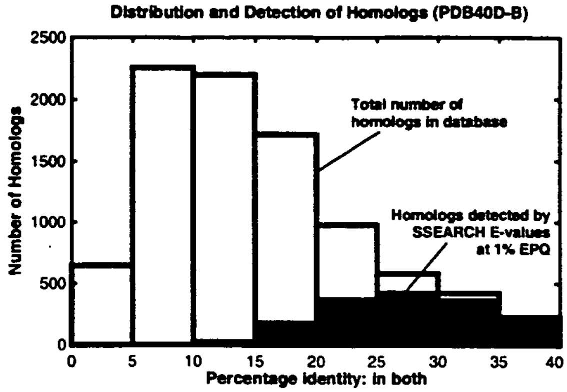


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.